Computational modelling of the relationship between *Miscanthus* genotype, phenotype and environment

Michael Graham Squance

Thesis submitted in fulfilment of the requirements for the degree of PhD.

Institute of Biological, Environmental and Rural Sciences

30th September 2014

DECLARATION

This work has not previously been accepted in substance for any degree and is not being concurrently submitted in candidature for any degree.

Signed ..... (candidate)

Date .....

STATEMENT 1

This thesis is the result of my own investigations, except where otherwise stated. Where *correction services have been used, the extent and nature of the correction is clearly marked in a footnote(s).

Other sources are acknowledged by footnotes giving explicit references. A bibliography is appended.

Signed ..... (candidate)

Date .....

[*this refers to the extent to which the text has been corrected by others]

STATEMENT 2

I hereby give consent for my thesis, if accepted, to be available for photocopying and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed ..... (candidate)

Date .....

Abstract

Several major global challenges being faced in the 21st century, ranging from climate change, energy security and food security to the sustainable living. Innovative solutions are needed to address those challenges. *Miscanthus* is a highly productive C4 grass which naturally occurs in Asia with the potential use for as a bioenergy crop. Recent advances in technologies such as genomics, phenomics, bioinformatics and modelling, provide a unique opportunity to accelerate the domestication process of *Miscanthus*.

Modern breeding programmes aim to utilise genetic information to assist in breeding decisions. High-throughput technologies such as genotyping-by-sequencing (GBS) generate massive datasets. Conventional analysis methods cannot handle large multi-dimensional datasets, therefore new methodologies are needed.

This research aims to use machine learning to model marker trait association and genotype by environmental interaction on *Miscanthus*. Three studies were performed in this research: 1) Develop a machine learning based QTL analysis tool to detect QTL on a *Miscanthus* flowering time mapping population. 2) Conduct marker-trait associations in a GBS analysis. 3) Establish a predictive model to understand drought and thermal effects on flowering time in *Miscanthus*.

The machine learning algorithm, random forest, was used to develop a QTL analysis tool, referred to as RFQTL. RFQTL identified several flowering QTL, with reduced computation time, consistent with conventional QTL analysis. Within the GBS study machine learning detected markers which when aligned with the *Sorghum* genome several homolog QTLs were found for the traits investigated. Using the prediction model of flowering time we were able to show that drought delays flowering whereas increased temperature led to earlier flowering.

This research has demonstrated the power of machine learning as an effective method for marker trait association and genotype by environment modelling. It has great potential to play a crucial role in crop improvement and provide further scientific insights for genetic research.

Acknowledgements

During the four years of my studies I have had the help and support of many people, here I wish to thank them.

Firstly I would like to express my thanks to my supervisors, Lin Huang, Iain Donnison and Ross King. Their help and support has been greatly appreciated over the past four years. I also wish to thank those at Ceres inc. who have provided advice and support, Richard Flavell, Tim Swaller and Xuefeng Ma. Nickolai Alexandrov and Charlie Rodger both formally of Ceres also receive my thanks for their support.

I wish to thank several members of staff from the IBERS *Miscanthus* breeding team, firstly John Clifton-Brown for his support and advice. I would like to especially thank all those who have helped me with phenotypical measurements and also have given endless help and support, Charlotte Hayes, Sue Youell and Maurice Hinton Jones. Special thanks should go to Richard Webster and Marc Loosley not only for providing help and support but also helping making the trips to the German field trials which are hard weeks not only enjoyable but also fun. Special thanks must also go to John Norris for his endless help with setting up servers, software and computing advice, helping with access to the department database and more importantly for providing advice and support over the years.

I also wish to pay thanks to many members of the IBERS environmental impact theme. Firstly I wish to thank both Elaine Jensen and Sarah Purdy for providing me with data and advice to use in my studies. Thanks to Gancho Slavov, Chris Davey, Paul Robson and Kerrie Farrar also for there advice over the years. I wish also to thank Astley Hastings of Aberdeen University for his expertise in processing and visualising climate data.

I would also like to give the greatest thanks to my partner Naomi Cope-Selby for her love and support, she helped me make it through to the end. I also want to thank my family for their love and support over not just through my PhD studies but my whole life.

Abbreviations

AFLP – Amplified fragment length polymorphism

ANN – Artificial neural network

GBS – Genotyping-by-sequencing

GS – Genomic selection

LAI – Leaf area index

MAS – Marker-assisted selection

MDS – Multidimensionality scaling

NGS – Next generation sequencing

PAR – Photosynthetically active radiation

PCA – Principal component analysis

RUE – Radiation use efficency

SNP – Single nucleotide polymorphism

SVM – Support vector machines

QTL – Quantitative trait loci

Table of Contents

List of Figures

List of Tables

1 Introduction

Since the industrial revolution, fossil fuel has become the major energy source for mankind. Fuel, fertilizers, solvents and pharmaceuticals compounds are just a handful of many products derived from fossil fuels. With dwindling resources of fossil fuel reserves and the realisation of increasing CO_2 and other greenhouse gas emissions (GHG) associated with climate change has driven UK and other nations to set an extremely challenging target of GHG emission reduction targets and has developed new policies targeting energy security and GHG mitigation (Karl *et al*., 2003; Crowley, 2010; Department of Energy and Climate Change, 2011). The UK government has set a target of 15% renewable energy production to be met by 2020. Tidal, nuclear, wind (on and offshore), solar and bio-energy are all touted as possible solutions. However, it is unlikely that single one of these will be able to meet the 15% target alone.

A more likely scenario for future energy production is each possible alternative fuel source will provide a portion of the total energy requirement. This will not only increase the robustness of future energy production but also allow achievable targets for each energy sector. Various studies have looked into what technologies should be used and how to build the best mix to provide a robust energy system from renewable sources (Pacala & Socolow, 2004; Bajpai & Dash, 2012; Erdinc & Uzunoglu, 2012).

Pacala and Socolow presented this mix as 15 possible stabilisation ’wedges’; each one contributing to a proportion of the CO_2 reduction. This concept not only addresses the idea of diverting the fuel sources from fossil fuels, but also to reduce the energy use through decreasing the amount of car usage and improvements in energy efficiency.

Plants are recognised to play a pivotal role in this mix. Through photosynthesis, plants

remove CO_2 from the atmosphere and convert it into dry matter (carbohydrate). This biomass can be used as a renewable feedstock for conversion into bioenergy (electricity and heat), biofuels (transport fuels) and biomaterials, thereby offsetting GHG emissions associated with fossil fuel usage and providing alternative sources of energy and products. The potential contribution of plants is not limited to fossil-fuel substitution, they are also the principal source of soil organic carbon, through below ground (via roots) and surface (via plant residues) inputs, and are the primary route to carbon sequestration in the terrestrial pool. Therefore, through maximising fossil-fuel substitution and carbon sequestration, land-based solutions can help to combat energy security and climate change (Lemus & Lal, 2005).

1.1 Global Energy Challenges and Bio-energy

There are several ways to address the renewable energy production. Wind, solar and tidal are some of the possible options to generate electricity but only the plant-based renewable energy is capable of providing the base chemicals needed for everyday items. Many studies have looked at using an array of biomass sourced platform chemicals to substitute transport fuels such as ethanol from biomass (Dodds & Gross, 2007; Bai *et al*., 2010; Farrar *et al*., 2011; Cherubini & Strømman, 2012). Bio-energy will play an important role in the future either as a fuel source or a source for platform chemicals derived from fossil fuels. Thus there is a real imperative to drive rapid innovative solutions from biosciences towards the goals of energy security and climate change mitigation.

The first group of plant species, namely first generation bio-energy, used to generate bio-energy came from food crops such as maize (de Vreis *et al.,* 2010). The grains produced by these species are often high in starch which was converted to sugars and

then to ethanol via fermentation. Some species such as Sugarcane (*Saccharum officinarum L*) are high in sugars that could be fermented for bioethanol production. Oil of oilseed rape (*Brassica napus L.*) can be extracted for biodiesel. One major concern of first generation biofuel was that they will compete with their food use and driving up the cost of food. Although studies suggested that food prices are more likely linked to the cost of crude oil rather than to bio-ethanol production (Flavell *et al*., 2011).

More recently, a second generation of bio-energy groups have emerged. They are dedicated bio-energy crops and are predicted to make a significant contribution to the future renewable energy mix (Somerville *et al*., 2010; Valentine *et al*., 2012). The majority of second generation bio-energy selected are perennial species due to their high efficiency in nutrient recycling. These include grass species such as *Miscanthus*, switchgrass (*Panicum virgatum* L.), reed canary grass (*Phalaris arundinacea* L.), and deciduous trees such as willow (S*alix* spp.) and poplar (*Populus* spp.). Being a perennial species, it reduces the need for replanted for many years, therefore lowering the total costs. However, an up-front payment would be needed for establishment. Some bio-energy crops such as *Miscanthus* have the potential to sequester carbon, which will help to mitigate the effects of CO_2 emission and derive the platform chemicals. With the expecting increases in use of bioethanol (OECD-FAO, 2013), it is clear that second generation bio-energy will play a major role in future energy and platform chemical production and carbon sequestration.

The EU 2020 directive for renewable energy requires the UK to generate 15% of its total energy from renewable sources (European Parliament and Council of the European Union, 2009). A DEFRA census in 2010 reported that currently 1.8% of the total land is used for bio-energy (Defra 2013). Within the UK *Miscanthus*, willow and waste products

from food crops, such as barley straw, are used to generate bio-energy. The March 2013 report produced by the Biomass Energy Centre has indicated that in 2013 a total of 1,090.2 MW_e was generated from biomass power stations. It also listed proposed and in planning biomass power stations (Biomass Energy Centre, 2013).

1.2 Breeding Miscanthus as Energy Crop

The most important plant species for biomass production are those with low inputs in terms of nutrients, husbandry and water, while producing high outputs in terms of yield. Also, it is expected that the crop species are selected to target specific climates and conversion processes (Flavell *et al*., 2011). The energy grass *Miscanthus* is unusual in being a highly productive tropical grass using the more efficient C4 photosynthetic pathway which is adapted across a very wide geographic region from the tropics in SE Asia through to Siberia (Vermerris*,* 2008; Heaton *et al*., 2010; Jørgensen, 2011). Generally, *Miscanthus* has been classified into 3 species: *M. floridulus, M. sinensis, M. sacchariflorus* (Chou, 2009; Hodkinson *et al.,* 2002a).

One naturally occurring species, *Miscanthus x giganteus*, a hybrid of *M. sinensis* and *M. sacchariflorus* (Hodkinson *et al*., 2002b), has been cultivated as a biomass crop in recent decades and although it is productive in temperate climate it is not highly tolerant to a number of stresses (Ings *et al*., 2013). It is expensive to establish as it can only be propagated vegetatively. Seed-based varieties are needed to reduce the establishment costs.

Given the variation of yield and other traits at different locations (Lewandowski *et al*., 2000; Gonza *et al*., 2001), it is unlikely that one single cultivar of *Miscanthus* will be used for all potential sites. Diverse *Miscanthus* cultivars should be developed to suit different

environmental conditions. The idea of having many diverse cultivators has been suggested in other plant species for redundancy against disease and pests (Finckh *et al*., 2000; Tooker & Frank, 2012). Also as future climate models predict extreme weather events are becoming more frequent (Meehl *et al*., 2000; Rosenzweig *et al*., 2001) along with the reduction in water availability (Olesen & Bindi, 2002; Schröter *et al.*, 2005) suggested that any new *Miscanthus* variety needs to have the quality to cope with changing environment.

Starting from 2004, a *Miscanthus* breeding programme has been set up at IBERS. The goals of this breeding programme range from the optimisation of crop performance to the understanding of its chemical compounds for conversion. It can then be processed and converted to a range of end products from energy through to chemicals and materials replacing a wider range of products currently manufactured using fossil and scarce resources. Since 2006, IBERS has taken a collection of diverse *Miscanthus* germplasm from the wild in Asia. Through the evaluation of performance in the UK and Europe, a number of technologies including genetics, modelling and bioinformatics etc. have been applied to dramatically reduce the cost of establishment of *Miscanthus* as an energy crop and to maintain and increase the natural genetic diversity of the crop.

In 2011, 8000 ha of land was used to grow *Miscanthus* in the UK (0.17% of total arable land) (Defra, 2013) with a total of 40,580 tonnes used in power stations between 2010 and 2011, which is approximately 2.5 times more than the year before. The 40,580 tonnes only accounts for less than half of the grown *Miscanthus* based on DEFRA's lowest estimates. From the 2010 agricultural census, it is estimated around 400 *Miscanthus* growers exist in the UK. The data from the DEFRA census showed an increase in the amount of *Miscanthus* used in power stations from 2009 to 2010. Studies have suggested the high cost of establishment could be the barrier to the uptake of *Miscanthus* by UK

farmers and suggest that establishment grants my overcome the barrier (Sherrington *et al*., 2010).

1.3 Quantitative Genetics and Marker-Assisted Selection for Molecular Breeding

Quantitative genetics is traditionally described as the study of genetic and environmental basis of the variation. Classical genetics typically deals with single genes of large effect, while quantitative genetics often investigate all genes as a whole and the total variation observed in a population results from the combined effects of genetics and environmental factors. It aims to predict the response to the selection by analysing phenotype data and relationships of individuals. One study theorised that if science could understand the effects of all genes, a breeder could 'cherry pick' those which could give the greatest advantage (Bernardo, 2001). However with current technology and understanding of quantative genetics this is currently impossible. Instead methodologies such as phenotypic selection, marker-assisted selection, genomic selection and genome wide association studies are used to facilitate breeding.

In conventional plant breeding, genetic variation is identified by visual selection. By selecting the best parents to create new generations of better performing progeny, this process is known as phenotypic selection (Kingsolver *et al*., 2001). However, with the development of molecular biology, variations in DNA can now be identified and studied for their effect on phenotype. Genetic markers have been widely used to facilitate studies of inheritance and the genetic variation of an individual, gene or cell (Mohan *et al*., 1997; Madhumati, 2014).

Recent advances in genomics, such as next generation sequencing and high-

throughput detection of single nucleotide polymorphisms (SNP), means that high density marker datasets can be generated economically (Rafalski, 2002; Elshire *et al*., 2011). Coupled with the fact that low-cost genotyping is much easier to achieve, approaches such as marker-assisted selection (MAS), Genome Wide Association Study (GWAS) and genomic selection (GS) have been used to drive the process of crop improvement (Jannink *et al*., 2010; Hamblin *et al*., 2011).

These improvements in genotyping technology mean that the generation of genetic data is often more economical than phenotyping (Bernardo, 2008). This potentially allows for the use of genetics in breeding, which is appealing as it could provide both cost and time saving. However phenotyping is still required for all approaches as it is required to develop models of the relationships between genotype and phenotype. Recent studies have shown developments of methods which aim to increase the throughput rate of phenotypical observations (Montes *et al*., 2007; Furbank & Tester, 2011).

Marker-Assisted Selection (MAS) is a process where genetic markers are used to speed up the selection process in breeding. The breeder can take advantage of the association between agronomic traits and allelic variants of genetic markers. Quantitative Trait Loci (QTLs)-based MAS, is one of the most widely used methods to detect genetic variances which affect phenotypic traits and to reduce the time needed to develop improved progeny within a breeding programme (Francia *et al.,* 2005). It has been applied in the breeding programmes of many crop species (Prasanna *et al*., 2010; Limure *et al.,* 2011; Steele *et al*., 2013; Ashraf & Foolad, 2013).

The simplest form of QTL analysis is to study single markers and calculate the probabilities of each marker's effects on the trait. However, this method is susceptible to genotyping errors and is unable to account for interactions between marker associated

QTL. Multiple-markers approaches are recognised to be more effective (Knott *et al.,* 1996).

QTL mapping usually involves creating, genotyping and phenotype mapping populations, generating genetic linkage maps, and establishing marker-trait association. The number of progeny is important when mapping QTL, since larger families have a greater potential for recombination leading to a high discovery rate (Darvasi *et al.*, 1993). Nonetheless, some literature suggests that little improvement is seen in populations over 300 genotypes (Vales *et al.*, 2005). Studies often detect a limited number of QTLs with large effects, although it is likely that many small effect QTLs exist for a trait that go undetected (Buckler *et al.*, 2009).

The efficiency and success of MAS depends on many factors associated with how the underlying marker and trait associations were identified. To name a few of these factors, they include the size of the mapping population, the nature and quality of phenotyping, the location of the markers with respect to gene of interest, the design and analysis of experiment, the number of markers available and the genomic region containing the desired QTLs etc. Another factors which can effect the efficiency of MAS is genotype by environment interactions. This effect could be accounted for by including environmental cofactors into the models developed, although this requires the mapping family to be replicated into several different environments.

As conventional breeding attempts to combine more target traits, there tend to be an overall loss of breeding gain and an increase in the duration of breeding cycle. Therefore, MAS offers great potential to improve the overall pace and precision of the breeding process by assembling target traits into the genotype more precisely and thereby reduce the breeding cycle.

Genome-Wide Association Mapping (GWAS) is another approach of analysing many

common genetic variants in different individual to see if any variant is associated with particular traits (Visscher *et al*., 2012). GWAS typically focus on associations between single nucleotide polymorphisms (SNPs) across different genotypes and their associated traits.

Normally, GWAS analysis is performed using mixed models (Yu *et al.*, 2006). One widely used tool for GWAS is the Genome Association and Prediction Integrated Tool (GAPIT) (Lipka *et al*., 2012). These methods require the construction of a kinship matrix to describe the relation of a genotype in the population, as population stratification can effect GWAS results (Ma *et al.,* 2012a). The information is then used in the modelling process to detect markers that are associated with the trait of interest. Many GWAS studies have been carried out to study the genetic causes of disease in Humans (S*cott et* al., 2007; Welcome Trust, 2007), with several possible SNPs have shown to have an effect on disease likelihood, but most of these studies can only explain a proportion of the effects (Manolio *et al*., 2009). The portion which cannot be explained if often referred to as the missing heritability. GWAS studies were also carried out on several crop species including rice (*Oryza sativa*), bread wh*eat (Triticum aes*tivum L.), Lolium and maize (Skøt *et al.*, 2005; Neumann *et al.*, 2010; Tian *et al.*, 2011; Zhao *et al*. 2011). However, same as the studies in humans, GWAS in plants is still unable to account for all the variance observed from phenotypic observations (Brachi *et al.,* 2011).

Genomic selection (GS) aims to use whole genome markers associating genotype with phenotype to inform breeding (Meuwissen, 2001; Jannink *et al.,* 2010; Ogutu *et al.,* 2012). In GS, a large number of markers are required for modelling the relationship between multi-genotypes and phenotypic values of targeted traits. The genome wide markers developed for GWAS can be used directly for GS study. Several studies have

indicated that for traits with a high heritability are more likely to be better predicted by GS, although exceptions do exist where traits with low heritability are able to be predicted with a high level of accuracy (Combs & Bernardo 2013; Luan *et al.* 2009). More recently, GS has been applied to crop breeding programmes (Heffner *et al*., 2009; Heffner *et al*., 2010; Sorrells *et al*., 2011).The idea of applying GS to crop breeding is due to the poor performance of MAS, where there are limitations of bi-parental mating and the limited power of current statistical tools for analysis (Heffner et *al.,* 2009).

Another study investigating the effectiveness of GS on wheat (Sorrells *et al*., 2011) has compared the GS with phenotypic selection (PS) and marker assisted selection (MAS). The study concluded that for all traits investigated, GS demonstrated an improved accuracy compared to MAS, and PS has similar accuracy with GS. The main appeal for GS is that the time between breeding cycles could potentially be reduced as prediction models would allow for estimation of gain without the need for years of evaluation.

To date only a small number of studies have been conducted in associating *Miscanthus* genetics with phenotype. Five QTL studies have been carried out and several of them were performed on the same family (Atienza *et al* 2003a, 2003b, 2003c, 2003d; Gifford *et al*., 2014). Others had studied genome wide associations in *Miscanthus* (Slavov *et al.,* 2014). All these findings had revealed and confirmed that *Miscanthus* has high levels of synteny with *Sorghum* (Swaminathan *et al*., 2010; Ma *et al*. 2012b). It is recognised that most complex traits are controlled by many polymorphisms with small effect (Buckler *et al*., 2009). Since the yield associated traits of *Miscanthus* are highly polygenic (Robson *et al.,* 2013), GS has the advantage of being able to deliver superior trait predictions of a polygenic nature.

An increased number of SNPs are becoming available because of new genotyping

technologies such as Genotyping by Sequencing (GBS) (Elshire *et al.,* 2011). Thus, analytical capability and computational time have become more of an issue when implementing MAS, GWAS and GS for molecular breeding. It is therefore one of the main objectives of this research to address these issues through the application of novel machine learning approaches.

1.4 Machine Learning as a Powerful Tool for Data-driven Biology

'Big data' refers to data acquisition methods that are considered to be high dimensional. The business sector has expressed great interest in 'big data' as credit card activity, website logs and data tracking methods have provided high volumes of complex data in which potential patterns about people's purchasing habits could be extracted and exploited for commercial gain.

Biological discovery, in general, has evolved considerably in the past two decades and increasingly being driven by the advances in new technologies, such as next generation sequencing (NGS), high-throughput molecular markers generation and genotyping and phenomics and other molecular tools that generate 'Big Data' (Marx, 2013). These high-throughput data generation methods tend to create high dimensional datasets where the attributes generated for a particular observation are much more than the number of observations.

Classical statistical inference cannot handle high dimensional data sets (Hastie *et al*., 2008). Hastie *et al* have pointed out that alternative approaches are needed to deal with high dimensional data. Machine learning (Mitchell, 1997) is a subset of artificial intelligence. It is a process of using algorithmic models to 'teach' a computer to understand a problem. The goal is for the algorithm to learn the complex patterns that exist within high

dimensional data to allow the prediction of future events.

Machine learning (algorithmic model) differs from the statistical approach (data model) which uses predefined assumptions of data distribution, to which data is then fitted in order to estimate the response. The algorithmic approach builds a model that treats the data domain as unknown and explains the output using the input by searching through a hypothesis space for the one that best fits the current problem. Both approaches can be utilised to provide understanding and prediction of complex system, however machine learning approaches via methods such as attribute subset selection, are more favourable for high dimensional problems (Hastie *et al.,* 2008). Therefore the machine learning approach could be more effective in the era of data-driven biology where datasets are massive, complex and multi-dimensional.

A study that investigated both statistical and machine learning approaches (Breimen, 2011b) have showed that the statistical approach was unable to provide satisfactory analysis, whereas, machine learning was able to effectively model the same problem satisfactory. However, Breimen also pointed out that this does not mean we should abandon the data modelling approach, but instead, the algorithmic approach can provide alternative tools for data analysis on high dimensional data sets.

Machine learning has been applied to many high dimensional problems in biology. Heslot *et al* looked at the application of a wide range of machine learning approaches for genomic selection (Heslot *et al.,* 2012). Several machine learning approaches were demonstrated to perform well for genomic selection. Other studies have looked at genotype identification from metabolites using machine learning and concluded that machine learning outperformed statistical analysis (Taylor *et al.,* 2002; Scott *et al.,* 2010). Both studies demonstrated the ability of machine learning to outperform statistical analysis

for metabolome analysis. The problem domain of these studies was high-dimensional with large numbers of attributes with limited numbers of observations.

Machine learning has also been applied to a wide range of problems that relate to the improvement of crops. One such application is the identification of traits which play an important role in yield. A study of maize traits revealed that many traits such as sowing date and soil type were important factors for yield (Shekoofa *et al.,* 2014).

Bernardo reviewed the molecular marker usage in crop improvement for the past 20 years and has confirmed that machine learning is one of several promising methods which can assist in developing new varieties (Bernardo, 2008). Bernardo suggests many breeders may have unwittingly used machine learning when performing chemometric analysis. However he also indicated that these methods are untested for predicting marker performance and further study is needed.

Another study aimed to use machine learning to classify a set of maize genotypes into heterotic groups using molecular data (Ornella & Tapia, 2010). The authors use support vector machines, bayesian methods and linear regression to perform classification. Bayes nets were shown to perform best for classifying heterotic groups.

Much published literature (Ornella & Tapia, 2010; Taylor *et al.,* 2002; Scott *et al.,* 2010; Shekoofa *et al.,* 2014) have demonstrated the power of machine learning for the analysis of massive volume of biological data. Its ability to handle high-dimensional data and to perform hypothesis discovery are of great use in the biological research and crop improvement.

1.5 Applying Computational Modelling to Predict Crop Performance under Different Environments

The ability to do accurate prediction is one of the most desirable goals of modern scientific research. Computational modelling is the use of methodologies from mathematics, physics and computer science to study the behaviour of complex systems and subsequently to be able to make informed predictions. The establishment of predictive models have been used in biology to understand and simplify complex systems. It has been used to predict future outcomes such as the effects of climate change on crop species (Summerfield *et al*., 1991; Lobell & Burke, 2010).

There are many justifications for the performing crop modelling. Hammer *et al* discussed this in their paper and suggested that crop modelling can be used for the development of heuristics techniques to inform scientific investigation and as a tool to understand genetic regulation to aid crop improvement (Hammer *et al.,* 2002). The authors highlighted that crop modelling has been used as a tool to understand the behaviour of various plant organs. Modelling provided a way to link all the investigations that have been done upon single systems. Traditionally, there are two schools of theory for crop modelling, 'mechanistic' and 'empirical'. Mechanistic modelling concerns itself with explaining how a system behaves whereas an empirical model aims to predict what a system will do. However, the line between these two is blurred. Most modelling approaches apply the same curve fitting methodologies to problems. The authors draw attention to the many uses of plant modelling including, education, decision support and scientific enquiry. Another type of crop modelling discussed in this paper are genomic models. Genomic modelling is suggested by Hammer *et al.,* as being likely to be one of the most important tools for crop development. To date, crop modelling has been performed in both energy

and food crops to provide yield predictions.

Several studies have been performed to predict *Miscanthus* yield. MISCANFOR is a model used to predict *M* x *giganteus* yield potential under different environment conditions (Hastings *et al*., 2009a). It estimated that *Miscanthus* would be able to provide 12% of Europe's energy needs. The MISCANFOR model was also applied to a theoretical drought tolerant *Miscanthus* hybrid (Hastings *et al.,* 2009b). Hastings *et al* used future climate models to predict that a decrease in water availability will occur as a consequence of climate change; this will lead *M x giganteus* to have a diminished yield. However the drought tolerant hybrid, refer to by Hasting as hi-tech, was shown to maintain its yield under future climate scenarios. The scenario presented by this study also concluded that if the drought tolerant hybrid was grown on 10% arable land, across all european countries, this would account for 3.6% of 2005 EU27 primary energy consumption; the production of which will mitigate 4.0% of total CO_2 GHG emissions.

Another study looked at modelling *Miscanthus* yield using lower resolution data (Pogson, 2011). Building upon the work by Hastings *et al,* this study aimed to use easy to measure meteorological observations to predict potential *Miscanthus* yields. The new model used cloud cover and latitude measurements to predict yield. This model was not tested against field data, but was compared to the MISCANFOR model. The low resolution model was shown to have a 0.68 correlation in prediction when compared to the MISCANFOR results. Although the low resolution model does not perform as well for prediction, the measurements are much easier to attain than those used in MISCANFOR. This represents a classic trade off in modelling, accuracy versus the effort required to attain data.

Another study examined the development of a model similar to MISCANFOR but

instead aimed at predictions in the USA rather than in Europe (Miguez *et al.,* 2012). The model was validated against published yields from several locations. It was demonstrated to have a high degree of accuracy, although the model did slightly over estimate the yield. Miguez *et al*'s model showed that *M x giganteus* yield was affected mostly by rainfall and the moisture holding capacity of the soil.

A recent study demonstrated that the potential cause of higher yields seen in *M x giganteus*, when compared to its parent species, is the improved radiation use efficiency (RUE) (Davey *et al*, in preparation). All species have similar light interception, regardless of canopy morphology. All *Miscanthus* species in the study were shown to close their canopy very rapidly in the early stages of growth. This indicated that optimal light interception was achieved early in the growing season. Therefore the study concluded that the reason some genotypes have higher yield was related to the efficiency the plants convert radiation into biomass yield rather than its ability to capture light. The authors compared data from two sites, Aberystwyth on the west coast and Rothamsted located towards the east of the UK. At the Rothamsted site, yields were lower than predicted by the model and suggested this is likely to be caused by drought with diminishing RUE. Other factors, such as maturity, were also suggested by the authors as contributing factors.

Another widely modelled bio-energy crop is short rotation coppice willow (SRC Willow). Willows (*Salix spp.*) are perennial, and like *Miscanthus* are being studied as a low input bio-energy crop (Karp *et al.,* 2011). Karp *et al* showed a wide range of genetic diversity available in wild willow species, and note that through breeding the yield has been increasing over the past forty years.

Several models have been developed to help willow breeders to understand the

important traits contributed to the crop yields. A model was built to study the ability of Willow to intercept light (Cunniff & Cerasuolo, 2011). They have demonstrated how clumping played an important role in the ability of different willow varieties to intercept light. Another study of light interception used pseudo 3D modelling to understand how the differences in the genotypes related to light interception (Cerasuolo *et al.,* 2013). These results from the model have indicated that the difference in leaf angles at various heights of the plants aided in light interception. The study also demonstrated that willow was able to rapidly adapt to an environment.

Genomic modelling through the development of a large number of mapping families and QTL mapping studies has played an important role in the breeding of new willow varieties (Hanley & Karp, 2013). They discussed the approaches used to develop genomic models for Willow. Fourteen mapping families have been created for willow, seven of which have been genetic mapped with one more in progress at the time of the publication. The willow breeding programme used these genomic resources and have discovered a large number of QTL in the K8 mapping population. The authors pointed out that in this modern world of genomics it is easy to forget the importance of phenotyping which is an important component of data collection for crop modelling. The willow breeding programme used data mining technology coupled with a central database and tools such as the Ondex software (Köhler *et al.,* 2006) for data integration. It is one of the prime examples of using genomics and modelling as powerful tools for both decision making and crop improvement.

Many crop modelling studies use classical statistical inference, however more recently machine learning has been employed in crop modelling. Machine learning was used to model the effects of water availability through irrigation and rainfall on the yield of

mango fruit (Fukuda *et al.,* 2013). Random forest was used in modelling to discover the best irrigation strategies for increasing yield of the fruit. Another study utilised the support vector machines (SVM) to predict brown rice yield based on meteorological information and nutrient availability (Saruta *et al.,* 2013).

Computational modelling have also been used to perform genomic selection (GS). Studies in wheat used several machine learning based algorithms, such as random forest and bayesian learning to build models. Heslot *et al* (2012) have compared several different models using multiple datasets for GS. The authors suggest that the non-linear nature of many machine learning algorithms may provide a better prediction to account for interactions that cannot be explained by conventional approaches.

It is clear that modelling has been used in a wide range of applications in crop improvement. These include the prediction of yields, understanding of light interception and more recently genomic models for breeding. Modelling has gradually become a crucial tool for crop science research and machine learning has been suggested as a prime candidate method for crop modelling. Further research of applying machine learning for crop modelling is needed to prove its effectiveness as a tool in crop science research and breeding.

1.6 Objective of the Research

The main objective of this research is to use the machine learning approach to analyse trait marker associations and to model genotype by environment interactions. Machine learning will be used to build predictive models as decision support tools for a genetically driven breeding programme. Machine learning methodology is also used to develop analysis tools for marker trait associations in both QTL and genome wide studies

to support marker assisted selection and genomic selection in breeding. This research will also demonstrate that machine learning is a powerful tool for modelling environmental effects on important traits. Attention is focused on the undomesticated bio-energy crop *Miscanthus* in this research. Other crop species are used to validate the developed tools and models.

To demonstrate machine learning's powerful capability as a tool in modern breeding programmes, this research also aims to answer the following scientific questions:-

- How computational approaches underpin quantitative genetics research?
- Why a machine learning can potentially increase the power of prediction of crop modelling to facilitate breeding programme?
- Would a machine learning/data mining approach be an answer to the association of complex Genotype-by-Phenotype-by-Environment (GxPxE) interactions?
- Will a machine learning approach be a better alternative than statistics to dissect the complex trait and conduct high-throughput markers analysis?
- Why and how a computational approach can help to drive 21st century breeding programmes?

1.7 Structure of Thesis

This thesis contains six sections. The first part is the introduction. The background and rationale of this research, ranging from the global challenge to the application of novel solutions to address issues in the 21st century, are described. In particular, the main objectives and the scientific questions which aim to be answered in this research are also presented in Chapter 1. Chapter 2 reviews the current state of research on various methodologies and materials used. These state-of-the-art research surveys provide

technological guidelines for the development of machine learning based applications and models in the later chapters.

Based on the machine learning approach, an efficient and versatile tool for QTL analysis is developed and described in Chapter 3. The analysis results have been compared with results from conventional QTL analysis to validate new machine learning approach.

Chapter 4 is devoted to the application of a machine learning approach to analyse the markers generated from high-throughput genotyping by sequencing (GBS) technology. The resulting model is then employed to study marker-trait associations.

Chapter 5 described the use of machine learning to build a predictive model to investigate the genotype by environment (GxE) interaction. The established model is then employed to predict the effects of drought and climate change on flowering time in *Miscanthus*. This thesis is concluded in Chapter 6 with discussion to answer the scientific questions raised in the previous section and the scientific significance of this research. The relevant topics for further investigation and research are also presented in this chapter.

2 Materials and Methods

2.1 Machine Learning and Data Mining

2.1.1 Machine learning/data mining: from theory to application

For the last decade, biological discovery is increasingly being driven by high-throughput technology. High throughput data acquisition methods are leading to the generation of a broad range of massive datasets. These datasets usually consists of many attributes whose nature and relationships are highly complex, in which valuable patterns may exist. With the increasing power and lowering costs of high performance computing, machine learning algorithms are becoming more widely used in analysing multidimensional data.

The concept of having machines which think for themselves and are able to adapt has long been a goal of computer science. This has led to the development of many artificial intelligence algorithms that are now easily available for application to new problems. The fundamentals of machine learning are addressed in Mitchell's book 'Machine learning' (Mitchell, 1997). Domingos, in his paper (Domingos, 2012) discusses machine learning in great detail, explaining the goal to build a generalised model from numerous examples.

Machine learning algorithms can be split into two major classes, supervised and unsupervised learning (Sathya & Abraham, 2013). Supervised learning is similar to the teacher and student relationship, with the algorithm (student) being given examples by the user (teacher). The user supplies observations for both the inputs and output and the algorithm will then attempt to understand how the inputs lead to the output. By presenting the algorithm with many examples one can re-evaluate and adjust the model in order to

formulate the general principle underlying the data in question. Eventually the model will be able to handle new examples and predict the outcome based upon previously observed information.

On the other hand, unsupervised learning does not have a teacher. Instead the goal is to infer the nature of the problem without being provided with correct answers or error on its decisions. This also means that there is no way to validate a models inferences. So in unsupervised learning heuristics arguments must be applied to both the algorithm and to its evaluation. Unsupervised learning problems are more complex than supervised learning due to not having data with which to validate the models that are generated.

Both supervised and unsupervised learning do not have to be applied independently and in practice we use both types of learning together. Various algorithms have been created to facilitate both kinds of learning approaches; each type of algorithm comes with benefits and caveats that will have different implications depending upon the type of problem.

Many machine learning algorithms involve first training and validating a model against a given dataset, once a valid model is found this can then be applied to new data to provide predictions. The generated model is unlikely to be changed unless its performance begins to degrade, upon which a new model may be generated using the original and newly acquired data. One of the main concepts held in machine learning is Occam's Razor (Domingos 1999), which concludes that the simplest solution is often the best. Therefore a model that is less accurate but simpler in its application might be superior to a complex model. Of course this depends on the problem but is commonly used as a strategy to avoid overfitting. Overfitting is a problem in all types of modelling but features heavily in machine learning. An overfitted model is one that has been over-trained on a set of

examples such that model performs better on the training data but performs badly when presented with new data. To test for overfitting a validation subset should be created. This subset is created from all the available observations and must be randomly sampled without replacement. The rest of the data is then used to train the model then the subset is used to validate the model to verify it has not overfitted.

Machine learning is capable of performing both regression and classification analysis (Witten & Frank, 2000). The attributes supplied to the algorithms can take the form of continuous or categorical data. These attributes can be a mixture of the two data types. Continuous data is considered to be a range in which a data point can exist. This can be either within a specified or infinite range. Categorical data consists of a set of discrete classes. A data point usually can only belong to only one of the possible classes, with the exception of fuzzy logic in which a data point could have partial membership to several classes. This type of data may or may not be ordered so that one class is considered to have a higher or lower value than another.

Some algorithms such as ID3 (Quinlan, 1986) are only capable of performing one type of analysis. In the case of ID3 it is classification. However many machine learning algorithms can perform both types of analysis. Although at times a user may wish to apply an algorithm that only handles classification, but where the dependant variable is continuous. In these instances conversion must be applied.

There are many strategies for converting continuous data to categorical data (Frank & Witten, 1999; Bay, 2000). One can choose split points in the continuous range and assigning data to classes by using the cut off points. Another option is the creation of dummy variables, which represent the a class to which a set of continuous data points are assigned to. The dummy variables are represented as a binary, where only one variable is

'on' at a time.

Data is at the heart of machine learning and is the key to the performance of the model. Most machine learning algorithms will choose which attributes it wishes to use within a model, which allows the algorithm to select the most informative subset of data and subsequently be used to filter out noisy attributes and those which do not add information. This clarifies which are the influential attributes and reduces the amount of data needing to be collected in the future, saving time and reducing cost.

The need for model validation has already been mentioned as a tool to identify overfitting. Validation of models is one of the most important stages in the development of a machine learning model. A subset of training data must always be removed from the model in order to test the model on unseen data. This dataset is used to test the model before it is applied to real world data, in order to confirm its accuracy. An example would be a model that was trained on a set of genotype and phenotype observations. Part of this data would have been set aside before model selection, known as the validation dataset. Using the remaining data the model is selected and trained. If the model performs accurately on the validation dataset it can then be utilised on new data. For the example this could be the next generation of progeny in a breeding programme, where only genotypic data is available which is then utilised to predict the phenotype of the progeny.

Active learning is a recent concept in machine learning whereby an algorithm can request new data from the user (Settles 2010). Alternately an active learner may receive data for which some are unlabelled, and the active learning may request for it to be labelled. This reduces costs as only data which were selected by the active learner need to be labelled. The learner may employ some heuristic model which means it is only requesting labels which it believes will increase its knowledge. Active learning

methodologies have been created using many different machine learning algorithms (Schein & Ungar, 2007; Burbidge *et al.*, 2007) and has been utilised in many fields such as text classification, image classification and drug discovery (Tong & Koller, 2002; Warmuth *et al.,* 2003; Hoi *et al*., 2006).

By allowing a machine learning algorithm to obtain a cost for data it forces the active learning to optimise its learning potential. The cost could be a function of time and money needed to get the label of a particular observation. For example in a breeding program this might be the cost of growing and phenotyping a plant. The idea is that the active learning algorithm will only request data which it believes will be worth the cost involved. It will then keep requesting new information as it requires it, and continuously retrain itself allowing it to adapt to new data.

2.1.2 Machine learning methods

As mentioned previously many machine learning algorithms exist, from tree based methods that provide clear interpretability of the representation of data, to neural networks which utilise formula which emulate the behaviour of neurons with our own nervous systems. Each algorithm has its own way of handling and representing data, and the choice of algorithm for a given problem is not straightforward. Several statistical approaches and machine learning methods will be presented here and their strength and weaknesses discussed and compared.

In the book 'Elements of Statistical Learning' Hastie *et al* (2009) describe in great detail a vast array of methods and theories that surround the idea of 'learning' from data. The second edition has added discussions on high-dimensional problems and random forests. Many of the methods described here are featured within in the book which

provides further reading on the methods and their application.

Linear Regression

Linear regression forms the basis of most statistical methods. Linear regression and its many variations have been widely used in many applications including genetics (Ogutu *et al.,* 2012), plant modelling (Robson *et al.,* 2013) and language processing (Gao *et al.,* 2006).

The formula for a linear regression model is

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

Where

y is a vector of response variables with length n

X is an (**n** x **p**) matrix

n is the number of examples

p is the number of independent input variables

β is a vector of length n consisting of coefficients that will be fitted

ε is the error term, accounts for any variance not explained by the model

There are multiple strategies available for fitting a linear regression model. Described here is the least squares estimation used for fitting a linear equation which produces minimal squared residuals for the data set. The following formula is for least squares minimisation. In the approach coefficient β is selected to minimize the residual sum of squares.

$$RSS(\beta) = \sum_{i=1}^{N} (y_i - x_i^T \beta)^2$$

Model coefficients are estimated via the following formula:

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

with fitted values given by:

$$\hat{y} = \hat{\beta} X$$

Logistic Regression

There are many extensions to the classic least squares linear regression. Logistic regression, for example, is designed to work with binary classification data.

Logistic regression is used a lot in classification problems which result in two possible classes, for example in disease resistance studies, where a patient is resistant or not. Due to logistic regression being used for comparisons in this work, the definition has been provided below.

Logistic regression is used to model the posterior probabilities for K classes using a linear function of x. It ensures that they sum to one, and remain in the range [0, 1]. The model is represented as

$$\log \frac{Pr(G = K-1 | X = x)}{Pr(G = K | X = x)} = \beta_{(K-1)0} + \beta_{K-1}^T x$$

It is fit using maximum likelihood for N observations

$$l(\theta) = \sum_{i=1}^{N} \log p_{g_i}(x_i; \theta)$$

where

$$p_k(x_i; \theta) = Pr(G = k | X = x_i; \theta)$$

Logistic regression is often used in data analysis where the goal is to identify attributes which explain the output. Often this is repeated over various subsets of the attributes. Logistic regression has been utilised in a wide range of problems including, genome wide associations (Wu *et al.,* 2009), microarray classification (Zhu *et al*., 2004) and spam filtering (Chang *et al.,* 2008).

Ridge Regression

Ridge regression (Hoerl & Kennard 1970), is a modification to linear regression that introduces a penalty to the sum of squares.

$$\hat{\beta^{ridge}} = \underset{\%bet}{argmin}\{\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=i}^{P} x_{ij}\beta_j)^2 + \lambda\sum_{j=1}^{p}\beta_j^2\}$$

λ is a complexity parameter which is greater than or equal to zero. It controls the amount of shrinkage, with a bigger λ value creating more shrinkage. The penalty takes the form of a sum of the squared values and is referred to as an L2 penalty. This can help to alleviate problems with correlated input variables. Input must be standardised prior to solving as ridge solutions are not equivariant under scaling. Scaling is not performed on the intercept. Ridge regression is based upon least squared regression, if we calculate least squares as

$$Y = \hat{\beta} X$$

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

Where

β is a vector of coefficients

Y is a vector of responses

X is matrix of input attributes

Ridge regression then modifies the formula as

$$\hat{\beta^{ridge}} = (X^T X + I\lambda)^{-1} X^T Y$$

where

I is a unit matrix with the same dimension as **X**

λ is the shrinkage parameter.

Ridge regression has been applied to marker assisted selection problems (Whittaker *et al.,* 2000; Ogutu *et al.,* 2012).

LASSO

The least absolute shrinkage and selection operator (LASSO) is another shrinkage method. It differs from ridge regression.

$$\hat{\beta^{lasso}} = \underset{\beta}{argmin}\{\frac{1}{2}\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda\sum_{j=1}^{p}|\beta_j|\}$$

There is similarity between this shrinkage function and that of ridge regression however the lasso penalty is a modulus whereas ridge regression is quadratic. This Manhatten norm penalty is referred to as an L1 penalty and is the sum of the absolute values. Given the nature of the constraint λ, attributes can have a coefficient of zero. This allows LASSO to perform a kind of attribute subset selection. LASSO has also been utilised in marker assisted selection (Ogutu *et al.,* 2012) and language modelling (Gao *et al.,* 2006).

Elastic Nets

Elastic nets is an algorithm that combines both the L1 and L2 penalty functions seen

in LASSO and ridge regression respectively. Elastic nets are used in genomic problems, where often strong correlations exist between genetic variables. LASSO is often indifferent to the selection between the correlated elements. Ridge regression however will shrink the coefficients towards each other. The elastic net penalty provides a compromise. The formula of the penalty is defined as

$$\sum_{j=1}^{p} (\alpha|\beta_j| + (1-\alpha)\beta_j^2)$$

Elastic nets have been applied in genomic selection and is one of the methods most frequently used (Croiseau *et al.*, 2011; Boichard *et al.,* 2012; Heslot *et al.,* 2012; Ogutu *et al.,* 2012).

Decision Tree Learning

Decision tree learning is a widely used machine learning method with several variations of the algorithm in existence. Decision trees work by partitioning the feature space. This partitioning creates a set of rectangles, into each a model is fitted. The simplest model being a single constant per partition. Decision trees can be used for both regression and classification.

In order to simplify the resulting model, only binary partitions are created. The feature space is split initially into two partitions. The split is performed on the attribute which will create the best fit. This is then repeated within each of the partitions. Each is again split into two. This continues until some stopping criteria is met.

Consider a regression problem, where the dataset consists of p attributes, with a single response and N observations. Assuming we have partitioned this data, using binary splitting, into M partitions, the response can be modelled as a constant c_m for each

partition R_M.

$$f(x)=\sum_{m=1}^{M} c_m I(x\in R_m)$$

If we use sum of squares for the criterion minimisation, $\hat{c}_m$ is just the average of y_i in the region R_m

$$\hat{c_m}=ave(y_i|x_i\in R_m)$$

Using minimum sum of squares to select binary partitions is computationally infeasible, therefore greedy strategies are often used. The following is an example of one greedy approach given in 'Elements of Statistical Learning' (Hastie *et al.,* 2009). Given j as the attribute to split and s as the split point, these can then be defined as a pair of half-planes

$$R_1(j,s)=\{X|X_j\leq s\} \quad \text{and} \quad R_2(j,s)=\{X|X_j>s\}$$

Then values for j and s need to be found which solve

$$\min_{j,s}[\min_{c_1}\sum_{x_i\in R_1(j,s)}(y_i-c_1)^2+\min_{c_2}\sum_{x_i\in R_2(j,s)}(y_i-c_2)^2]$$

For any choice of j and s, the inner minimisation can be solved by

$$\hat{c}_1=ave(y_i|x_i\in R_1(j,s)) \quad \text{and} \quad \hat{c}_2=ave(y_i|x_1\in R_2(j,s))$$

This algorithm can be performed much quicker than the minimum sum of squares so therefore it is feasible to scan all attributes for the best split point.

The next question is where to stop the tree growing. A simple method such as only allowing splits which decrease the error of the tree may be applied. However this might lead to informative splits being missed. For example, one split may not decrease the error but may lead to another split which does. A preferable strategy might be to grow a large

tree, stopping only when a predefined node size is reached, and then pruning the resulting tree. A variety of pruning algorithms exist that will remove branches from the tree. These branches may or may not increase accuracy of the tree, however their removal would create a simpler model, thereby satisfying Occam's Razor.

The application of decision trees to classification data uses different algorithms to decide split points and perform pruning, several examples of which are outlined in Hastie *et al* (2009). Decision trees have been applied to a multitude of problems including text parsing (Magerman, 1995**),** cancer detection from mass spectrometry data (Kaplan, 2003) and for disease diagnosis (Tanner *et al*., 2008).

Random Forest

Random forest is a variation of the decision tree learning with additional machine learning concepts. First presented by Breiman (Breiman, 2001a), random forest makes use of bagging (Breiman, 1996) and bootstrapping to improve the prediction power of decision trees. Bagging is a method that creates multiple predictors for a given problem, these are then grouped to form a single predictor. This provides either an average (regression) or consensus (classification) of all the predictors. Bootstrapping is the generation of multiple data sets by sampling with replacement from a training data set. These 'bootstrapped' data sets are then used to train the model. The model's fit is then examined across all the data sets.

As the name random forest implies one creates many trees to form a forest, with each tree being built using a different subset of the total data available. The number of trees can be specified by the user. In the randomForest library for R, this is done using the ntree parameter. A modification to traditional bootstrapping is used in random forest analysis.

Attributes are selected with replacement to form multiple datasets. The number of datasets is equal to the number of trees and then a tree is then created for each. The number of attributes within each bootstrapped dataset can also be user defined. This is specified using the mtry parameter in the randomForest library for R. However the following are the recommend values for mtry.

$$p/3 \; for \; regression$$

$$\sqrt{p} \; for \; classification$$

Predictions are taken as the average of all the trees,

$$\hat{f}(x') = \frac{1}{B} \sum_{b=1}^{B} \hat{f}_b(x')$$

where **B** is the number of trees

For classification the most often predicted class is selected from the trees in the forest as the answer.

The selection of a random subset of attributes for each tree means that it is unlikely that a single attribute would appear in every tree. This means that a single noisy attribute should not effect the model as a whole. This suggests that random forest is tolerant to noise. The bagging portion allows random forest to generalise easier as the prediction is in fact the consensus of several different weaker models instead of a single model. This again will help to reduce the effect of noise on overall result. Breiman's implementation of random forest algorithm is employed as randomForest in the R package of the same name (Liaw & Wiener, 2002). The random forest algorithm is an inherently parallel problem therefore it performs well using parallel computing. This parallelisation can be achieved using the SPRINT library (Hill *et al*. 2008) in R.

In order to understand what attributes are important within the model created by random forest an importance score can be calculated to provide a ranking of attributes. The following is the method used in Breiman's paper and the R package. Firstly the random forest model is created from a given data set. During this process each out-of-bag error of a data point is calculated, and then is averaged across all the trees in the forest. The values of each attribute are permuted from the training data, and the out-of-bag error is recalculated on the newly perturbed data set. The 'out-of-bag' error is calculated for each attribute by first creating a random forest predictor on all data. The prediction is calculated over the whole forest. A second prediction is calculated by averaging across all the trees in which the attribute does not feature. Then the error is calculated as the difference between the two predictions. The importance score is therefore calculated from the average of the difference seen in the out-of-bag error from before and after permutation. Attributes with a high score are considered to be the more important ones.

Random forests have been applied to genomic selection (GS) problems, (Heslot *et al*. 2012). Heslot *et al* has noted its effectiveness but warned that the method is unproven in GS and therefore should be utilised with caution. Another study showed the improvement in plant identification from metabolite fingerprinting by using random forest instead of principal component analysis (Scott *et al*. 2010). Other studies have made use of random forest in order to model fitness of DNA adaptors (Knight *et al*. 2009).

Artificial Neural Networks (ANN)

An artificial neural network (ANN) is an example of machine learning approach designed to mimic the real world and the working of a brain. It consists of an interconnected network of nodes, consisting of one of more hidden layers. These hidden layers connect the input layer to the output layer. The neuron component of the brain is

copied leading to a node that is 'fired' using the sigmoid function

$$\sigma(x)=\frac{1}{(1+\exp(-x))}$$

In order to fit a model weight are introduced to each node as such

$$\sigma(sx)=\frac{1}{(1+\exp(-sx))}$$

where **s** is the weight to be fitted. Weights allow each node to be adjusted so a model can be created. For a regression problem with only a single hidden layer sum-of-squared errors can then be used to fit a model.

$$R(\theta)=\sum_{k=1}^{K}\sum_{i=1}^{N}(y_{ik}-f_k(x_i))^2$$

where θ is the complete set of weights.

Another training algorithm for ANN is the back propagation algorithm (LeCun *et al*. 1989). It randomly sets all the weights in a predefined neural network. Training examples are then passed through the network and the weights are adjusted to minimise the error. To prevent the algorithm from constantly adjusting the weights, a learning rate is selected to limit the amount a weight can be adjusted for each observation. The lower the learning rate the longer it takes for the model to fit to the problem but it is less likely to overfit by falling in a local minimum in the problem space.

ANN's have been used for a range of problems including rainfall-runoff modelling (Dawson & Wilby, 1998), financial and economic predictions (Kaastra & Boyd, 1996) and algal bloom modelling and prediction (Recknagel *et al.,* 1997).

2.1.3 Statistical vs machine learning

Data analysis is a staple of all scientific pursuits, but especially in data driven science. There are several options for analysing data, these include statistical analysis and machine learning. Statistical analysis is concerned with fitting models to a predefined distribution. Machine learning, a sub field of artificial intelligence, provides an alternative. This has led to a question of which is better machine learning or statistical analysis?

There are a lot of similarities between both statistics and machine learning. Tibshiriani has summarised and produced a table the difference in terminology used in the two fields. Table 2.1 is a reproduction of the original table shown on Tibshiriani's webpage (http://statweb.stanford.edu/~tibs/stat315a/glossary.pdf).

Machine Learning	**Statistics**
Network, graphs	Model
Weights	Parameters
Learning	Fitting
Generalization	Test set performance
Supervised learning	Regression/classification
Unsupervised learning	Density estimation, clustering

Table 2.1: Table produced by Tibshiriani to illustrate the different terminology found in machine learning and statistics and their overlap

With the exception of the last two entries in the table, statistics and machine learning use different terminology to represent the same thing, and actually these terms are used interchangeably with a lot of the machine learning studies using the terminology found in statistics, such as regression and classification. Although they both share the same goal, they differ in the way they analyse data.

When it comes to data analysis there appears to be two main schools of thought, the data model approach, typically applied in statistics, and the algorithmic approach applied in machine learning (Breiman, 2011b). Breiman summarises this in his paper 'Statistical

modelling: the two cultures'. Breiman concludes that the main difference between the two approaches is in the way they treat the problem and link the inputs and outputs.

Statistical analysis is based on the data model, making prior assumptions about the distribution of data that it must fit. This could be a normal distribution with the goal being to discover the mean and standard deviation from the observed data. This is the case for methods such as linear regression and ANOVA, which assume all data is normally distributed and data points are independent of each other. Machine learning, on the other hand, is based on the algorithmic approach. It is not concerned with the distribution of data but instead looks for patterns existing between input variables and the response they produce. Although they differ by application both approaches aim to process a set of observations and then represent it in a formal way that can later be reapplied to provide understanding and prediction.

Breiman presented the following example where machine learning and statistics have been applied to the same data but produced different results. Breiman looked at the survival rate in 155 hepatitis patients. Two studies had previously analysed this data, the lowest error rate seen in these studies was 17%. Using the same dataset Brieman applied logistic regression and was able to reproduce the lower error rate. Random forest was then applied on the same dataset resulting in an error rate of 12.3%, a 30% reduction. Random forest revealed that most of the predictive power comes from two variables, 12 and 17. Logistic regression had suggested that variables 7 and 11 were the most important. However when variables 7 and 11 were modelled together using logistic regression the error rate was 22.9%. Whereas when 12 and 17 were modelled as single variables using logistic regression the errors where 15.7% and 19.3% respectively. This implies that the variables selected by random forest are more informative, and therefore it

can be concluded that the machine learning approach out performs the statistical approach for this example.

Several studies on classification of plant metabolites have made comparisons between machine learning and statistics. Taylor *et al* investigated the problem of identifying genotypes from their metabolome, and also discriminating between the progeny, which would only differ in the maternally inherited mitochondria and chloroplasts (Taylor *et al*., 2002). The authors concluded that linear discriminant analysis was ineffective for genotype discrimination. Instead an artificial neural network (ANN) was applied and was able to correctly classify 26 out of the 32 to distinguish between the parental types and the progeny, although it struggled to correctly classify the two classes within the F1 generation. The performance of the ANN was assessed using leave-one-out validation due to the low number of samples. This study demonstrates that machine learning can perform genotype discrimination using metabolome data whereas the statistical method failed.

A second study on Arabidopsis attempted to fingerprint mutants from their metabolome (Scott *et al.,* 2010). This was facilitated by using machine learning. Several methods, including principle component analysis (PCA), support vector machines (SVM) and random forest were used to fit the data. When it came to discriminating mutants the SVM's performed the best, but the other machine learning techniques performed equally as well, and in some cases better than the current fingerprinting approaches based on PCA. In Scott *et al*'s study the machine learning out-performed the statistical approaches. Although the SVM did outperform the random forest in this particular example, the authors note that the random forest did provide more interpretable results than the SVM model.

Although this only represents a small amount of the studies that compare machine

learning and statistics, it is clear that machine learning is just as powerful as statistics and in many cases outperforms it. It is also obvious that the two approaches are not opposing to each other but instead are tightly linked in terminology and goals. Each possesses different strengths and weaknesses. There is no golden hammer, i.e. a universal solution that can be applied to all problems.

Within the field of statistics there are many methods for analysis based on different data distributions and theories. Machine learning is also made up of many algorithms each leading to different data representations and each is loaded with different biases. The selection of a method therefore very much depends on the nature of the problem at hand. Machine learning is well known for its effectiveness in high dimensional problem spaces. With the recent 'big data' revolution does this mean that machine learning may come out the winner in the battle of data analysis?

2.1.4 Strength and power of machine learning

This section will examine the strengths of machine learning and discuss which methods are best suited for which type of problems.

Machine learning is the development and application of algorithms that allow a computer to learn without the need to be explicitly programmed. Machine learning is capable of analysing a variety of problems. Machine learning techniques have many real life applications such as, computer vision (Shafiee *et al.,*2014), search (Agichtein *et al.*, 2006), natural language processing (Daelemans & Hoste, 2002) and bioinformatics (Larranga, 2006). All these applications are complex systems, where the link between inputs and outputs are not clear due to the large number of possible attributes.

As mentioned previously the main difference between statistics and machine learning

is the way they fit different models to interpret data. The statistical approach makes assumptions about the distribution of data and then attempts to fit the data into this assumed model. However the data most likely will not satisfy all the assumptions, however the model created may still be of use, if it provides a good prediction. On the other hand, machine learning does not make assumptions about the data distribution. This means that machine learning attempts to learn the nature of the data and formulate the pattern that best links the inputs and the outputs. Therefore, machine learning is much better at handling complex problems and datasets without making the assumptions as statistics does.

Complex problems are often the high dimensional problems, in which

$$p >> N$$

where p is the number of attributes (or features) and N is the number of observations available for learning. Finding influential attributes in these high-dimensional data sets allows for creation of simpler models that would be useful for predicting or classification of new instances. In high-dimensional problems methods such as linear regression break down due to the complexity of the matrix algebra.

Many machine learning algorithms are noise tolerant. One recently developed machine learning method capable of dealing with noisy attributes is random forest. It makes use of bagging to allow for noise tolerance. Bagging generates multiple models from randomly selected data sets, this therefore prevents single attributes with large amounts of noise from effecting the overall model while also having the advantage of improving the ability of the 'bagged' model to provide a generic prediction. Other ways to deal with noise include implementing learning rates to allow a single instance to only have a small effect on the overall model. By minimising the effect of data points on the model

this means that an incorrect data point should not have a large effect on it. Data points whose values are similar will have a large cumulative effect over the training period than those which are likely outliers caused by noise. This strategy is used in fitting neural networks, a scaling factor or learning rate is applied in the back-propagation algorithm, which limits the amount a weight can be adjusted per observation (Widrow & Hoff, 1960). The choice of which method to use to handle noise is dependant on which algorithm you select as each have different strategies to deal with noise. However firstly some statistical analysis should be applied to attempt to understand the nature of the noise that exists in the data. A single noisy data point cannot have a great influence on the model. The goal is to understand the general pattern of all observations rather than allowing a single instance to largely effect the model. This generalised model is then tested against the validation data to ensure over fitting has not occurred. Some algorithms ability to deal with noise is traded off for other gains such as less storage requirements (Schlimmer & Granger 1986; Aha *et al*., 1991).

Computational biology is usually dealing with high-dimensional problems with tens of thousands of markers across hundreds of genotypes (Zhao *et al*., 2011; Ma *et al*., 2012b). In this application one is often interested in finding a signal for a gene, expression level or metabolite that effects a phenotype, classification or behaviour.

In two previously mentioned studies of metabolites (Taylor *et al*., 2002; Scott *et al*., 2010) the authors successfully applied machine learning techniques to detect useful signals in high-dimensional data and showed that they outperformed classical statistical methods for each of the applications.

There are many other application areas where machine learning techniques are used for complex data analysis. Sharma *et al* (2011) for example constructed a model to predict

solar power generation in 3 hours from the current weather information. They compare the results from machine learning approach with other models and show that the machine learning model performs better on the test set. The SVM learning had problems with redundant attributes, so PCA scores were used to remove the less influential ones. The SVM/PCA combination for attribute selection and prediction outperformed all other methods tried by the authors.

Another example investigated the use of machine learning approaches as a decision support tool in fraud detection in 49 previously published papers (Ngai *et al*., 2011). This paper addresses the many applications that utilise machine learning within the financial sector. Authors classify the types of analysis into six groups: outlier detection, clustering, classification, prediction, visualization and regression and demonstrate a wide range of algorithms that have been used, from decision trees and Bayesian methods to fuzzy logic.

The next study looked at the use of data mining in crime data (Chen *et al*., 2004). The authors highlighted several ways that machine learning is used to perform pattern detection for various types of data, structured and unstructured. They point out the use of data such as social networks to detect criminal organisation structure and the use of classification to detect linguistic patterns in spam emails to identify their source. They highlight the Coplink project (Chen *et al*., 2003) which applied machine learning to police reports and other data sets which were noisy and hard to process. The first example they presented was using a modified AI Entity Extraction to extract data from reports which contained typos and grammar mistakes. The AI Entity Extraction system uses a three step process, firstly it identifies noun phrases. It then calculates a feature set based on pattern matching and lexical lookup. Finally these features are then analysed using a neural network to predict the most likely entity type for each phrase. The AI methods were

capable of performing above average for the criminal's name and the name of the narcotics. However, it struggled to extract addresses and personal details. Another example was the use of clustering to model the structure of criminal organisations. Results were comparable to those obtained by human analysts.

The next application of machine learning moves away from the world of crime or energy to the world of archaeological artefacts; it tried to reconstruct frescos from fragments (Funkhouser *et al*., 2011). One problem faced when reconstructing artefacts is how to put the damaged pieces back together. In essence, the problem is a large jigsaw puzzle without a reference. Fragments were imaged and scanned to get the colours, shapes and 3d structure. Once data was acquired the 3D information was used to score the likelihood of every two pieces fitting next to each other. Once pairs had been identified their features were extracted and used to train an M5P decision tree using Weka. At each node the M5P tree fits a linear regression to the two subsets of the data created by the split. The models were trained on three different data sets and the authors concluded that machine learning provides an accurate method for the reconstruction of fresco pieces and by training on one fresco, others can be analysed using the same model.

Another study reviews the use of machine learning in the world of e-commerce data, where many companies are looking to leverage the vast amount of data they have collected on customers to increase their profit margins. One way to do this is to make sure advertising is targeted at the customers who are most interested in the product featured within the advert. The effectiveness of machine learning has been investigated for this purpose (Perlich *et al*., 2013). Advertising is a big business with billions of auctions for advertising space occurring daily. With data on customer behaviours, brand-orientation and the need for decisions to be made in a split second to win the best advertising slots,

complex data analysis must be exploited and this is where machine learning plays a pivotal role. The system illustrated in this paper used a two stage machine learning method. In the first stage the high dimensional sparse data were analysed, which is often marred by biases to identify the features to be used in the main learning task. The second stage used the selected features and weights and recalibration to learn the “target” distribution. The case studies have demonstrated that the machine learning method was able to increase sales and/or decrease the cost per download/registration.

High-dimensional problems are becoming more common in the modern world where high throughput methods, open access databases and automated data collection have resulted in generation of large data sets. These data sets have many complex interacting features where the most influential features must be extracted to form various hypothesis and provide cost effective predictions of future events. From the above discussion, it is clear that machine learning has many advantages over statistics when dealing with large data sets with unknown number of attributes hidden in a sea of other factors that need to be identified. Machine learning is also shown to be highly flexible, in that it can be used for many different types of problems and has been applied to a wide range of fields. In high dimensional problems where many factors are not influential in effecting the end results, machine learning will be the better choice.

2.2 Quantitative Genetics and Marker-Assisted Selection (MAS)

2.2.1 Quantitative genetics and molecular dissection of complex trait: theory and practice

The genome, the underlying code which holds the instructions for translating the many proteins that make up life, remains a complex mystery. Recent advances in genomic

sequencing and high-throughput methods have provided opportunity to quantify a genome through gene sequences and markers. This genomic revolution allows science to gain insights into biology and quantitative genetics with the aim of uncovering the links between genotypes and phenotypes we observe in an organism.

Quantitative genetics attempts to understand the nature of the genome through mathematical models in order to provide a link between observed phenotypes and the genes that underpin them. 'Introduction to Quantitative Genetics' (Falconer & Mackay, 1996) describes in great detail many of the concepts of quantitative genetics, in this section we briefly cover the core concepts.

First we will define two terms that are needed in quantitative genetics. Locus, which refers to a specific position within the genome. This can either be a region of DNA or a single base. The second term, allele, refers to variations at a given locus. If we take for example a single nucleotide polymorphism (SNP), this is a single base of DNA. Its locus could be for example on chromosome 1, at 1,567,843bp (base pair). If we assume the SNP is bi-allelic, meaning it has two possible values, i.e either C or T. We often recode these values to a standardised representation, such as A and a, these being the alleles of this SNP. A diploid organism would therefore have one of these three genotypes for this locus, AA, Aa and aa. The different alleles at this locus may potentially have an effect on the plants phenotype, if for example one of the alleles is linked with a damaged copy of a gene. Quantitative genetics is the study of alleles and how we can relate them to phenotypical differences.

In order to model the effects of an allele, we must first understand its motion through a population, and how this can then be represented mathematically. However before looking at the motion of a locus in a population we must first be able to define its current

allele frequency. If we assume a single locus that is bi-allelic in a diploid organism, its frequencies can be defined using the Hardy-Weinberg (HW) equation. Firstly assume an idealised population that is under no pressures from selection, migration or mutation. The following is the derivation of the HW equation.

Genotype	**AA**	**Aa**	**aa**
Frequency	**X**	**2Y**	**Z**

The various genotypes can be in any frequency as long as the following holds true

$$\mathbf{X} + 2\mathbf{Y} + \mathbf{Z} = 1$$

During reproduction each parent creates a gamete. A gamete is a sex cell, which contains only a single copy of each chromosome. Therefore a gamete will contain only one form of an allele, in our bi-allelic example either A or a. As we get two gametes, one from each parent, these will merge to make a zygote, recreating our alleles AA, Aa and aa. The frequencies of two gametes are calculated as such,

The frequency of gamete **A**, which will be represented with the symbol **p**, is given by

$$\mathbf{p} = \mathbf{X} + \tfrac{1}{2}\,\mathbf{2Y}$$

The frequency of gamete **a**, which will be represented with the symbol **q**, is given by

$$\mathbf{q} = \mathbf{Z} + \tfrac{1}{2}\,\mathbf{2Y}$$

And as the sum of frequencies of all the genotypes is 1, therefore,

$$\mathbf{p} + \mathbf{q} = 1$$

Then for the next generation the potential genotypes are calculated.

	Female Gamete	
Male Gamete	**A**	**a**
A	**AA**	**Aa**
a	**aA**	**aa**

Then by substituting in the **p** and **q** defined in the previous step

Genotype	**AA**	**Aa**	**aa**
Frequency	$\mathbf{p^2}$	$2\mathbf{pq}$	$\mathbf{q^2}$

Now in the new generation, the frequency of gamete **A** is

$$\mathbf{p}^2 + \tfrac{1}{2}\, 2\mathbf{pq} = \mathbf{p}(\mathbf{p}+\mathbf{q}) = \mathbf{p}$$

The ½ of the heterozygous frequency refers to the fact that from the Aa genotype there is a 50% chance of getting an A gamete. Therefore we see that under the assumption of random mating and no selection pressures or mutations the frequency of the alleles will remain the same from generation to generation. The assumption of random mating is not likely to exist in a real population for it assumes the infinite population size where there is no fitness or spatial selection.

The idea of an infinite or large population size means that the probability of an allele that has no selection pressure ever being lost through random mating is low (in an infinite population in fact no allele is ever lost unless selection occurs). Reduction of population size to a few hundred or thousand individuals means the chance of an allele being lost by random mating becomes much higher. Effective population sizes refers to the number of potential breeding individuals that exist in a population. The number of breeding individuals is the number of individuals that would exist in an idealised population that would result in the same variation of alleles under random genetic drift. These idealised populations are

usually much smaller than the number of individuals in the actual population. In humans, for example, it has been estimated to be 10,000 (Eller *et al*., 2011), and it is ranging from 32,500 in wild wheat to 12,000 in domesticated wheat (Thuillet *et al*, 2005). This means that the actual effective size of our population is smaller than the number of individuals.

In reality large populations are usually made up of smaller sub populations that may be caused by environmental factors or by distances between them. This is sometimes referred to as the island model where each population is considered to exist on its own island, separated by a large ocean, with either no movement between populations or with some amount of migration.

In small populations alleles can be lost under random mating and the smaller the population size the faster this can occur due to the other allele becoming fixed in the population. This is accelerated if selection is accounted for, which will be discussed later, but in reality loci can be recovered even when they are lost from a population. Two mechanisms exist in nature which can recover a lost allele. One is mutation and the other is migration.

Mutation is the change of a nucleotide within a genome which is not repaired. These can occur in both genomic and non-genomic regions. Mutation may lead to a change in phenotype, if it alters a gene. These alterations could be a change in the gene product, or the prevention of function. Mutation is a very rare event in the real world with mutation rates being as low as 2.2×10^{-9} per base pair per year in mammalian genomes (Kumar & Subramanian 2002), so its effect is highly limited unless something happens that affected the mutation rate. Let us assume that a mutation can occur in both directions, i.e. it is possible to change one allele into another and back again. If we considered a situation where u and v are the forward and backward mutation rates respectively. If we have two

alleles A_1 and A_2 each with an initial frequency of p_0 and q_0 respectively.

$$\begin{array}{ll} Mutation\,Rate & A_1 \underset{v}{\overset{u}{\Leftrightarrow}} A_2 \\ Initial\,Frequencies & p_0 \quad q_0 \end{array}$$

Then the change in allele frequency (Δq) in one generation can be found by

$$\Delta q = up_0 - vq_0$$

When it is assumed that small populations exists due to the island model as discussed above, then a second process exists for affecting allele frequencies and facilitating the recovery of lost alleles. Migration is movement of individuals from one 'island' to another. It can introduce new or lost alleles into a population or could just create a sudden shift in the frequencies of alleles within the population. The change of frequency from migration between the native populations (at frequency q_0) and the newly mixed population (q_1) is a function of the number of migrants. It is expressed as a proportion represented as m where q_m is the frequency of a certain allele among the migrants, such that

$$\begin{aligned} \Delta q &= q_1 - q_0 \\ &= m(q_m - q_0) \end{aligned}$$

When comparing the two effects, mutation is not as effective when migration is also present. Even a low rate of migration far outweighs the power of mutation within a small population, as long as the mutation rate remains low. So far we have assumed that the change in allele frequencies had no effect on the characteristics the plant displays. In reality different alleles could have an effect on phenotype. This change could potentially be advantageous to the organism, or could have a negative effect that might affect its ability to survive. Alternatively it could act to increase or decrease the chance of an organism reaching a stage of life to produce progeny. This change clearly will have an effect on the

frequency of an allele due to selection pressures which could be environmental, such as the ability to survive colder temperatures. Alternatively the change could affect the organism's ability to breed by reducing fertility or rendering it sterile.

The effect this has on gene frequency is known as selection pressure. Selection pressures are highly effective at adjusting allele frequencies and can cause rapid loss of an allele if there are no other forces effecting frequency, such as migration or mutation.

We refer to the contribution of the offspring to the next generation as its fitness. A fitter individual is more likely to survive and therefore breed. If this fitness is related to particular allele then selection operates upon it. The strength of a selection pressure against an allele is represented as the coefficient of selection, s.

Dominance is used to describe the different selection pressures based upon alleles. There are four allele dominance models which are used when discussing selection. They are represented graphically below for a bi-allelic diploid organism, with alleles A_1 and A_2. An additional term, h, refers to the fitness of the heterozygote. This only features in the partial dominance model, and is used to account for the different selection pressure seen on the heterozygote.

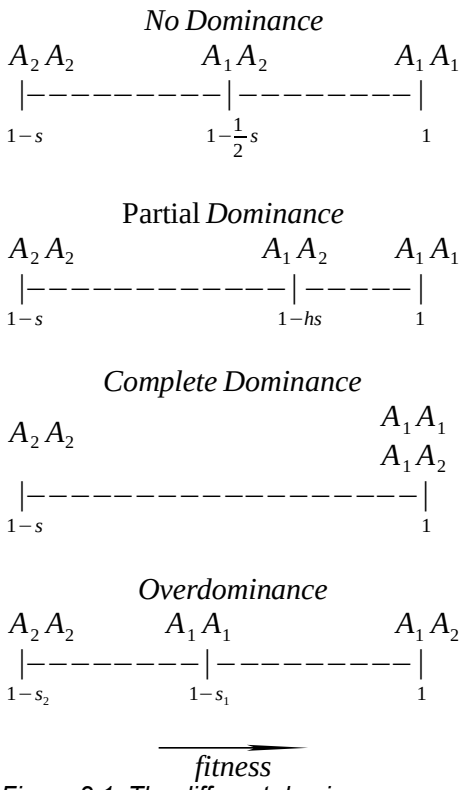

Figure 2.1: The different dominance models displayed by alleles, and there effects on fitness

We see in Figure 2.1 the four models of dominance. In all models the fittest allele is assumed to have a fitness of 1. The coefficient of selection then acts against the weaker alleles reducing their fitness. With the exception of the over dominance model one of the homozygotes is always considered to be the fittest, with selection acting against one or both of the other alleles. In the case of the first two models selection acts upon both the less fit homozygote and the heterozygote. However the heterozygote is either half as fit as the best allele or some proportion between the two homozygotes. In the complete dominance it is assumed that the presence of A_1 always grants a fitness advantage even when an A_2 is present, so the heterozygote also has a fitness of one. In the final model the two homozygotes are less fit than the heterozygote. In this case two selection pressures are needed s_1 and s_2 that act on each homozygotes.

The change in frequency of the alleles due to selection differs depending upon the effects of dominance. The comprehensive formulae can be found in Falconer's book “Introduction to Quantitative Genetics” (Falconer & Mackay,1996). Below is the formula for an allele with the frequency of q where the heterozygote displays no dominance. We assume this allele is under some selection pressure, s.

$$q_1 = \frac{q - \frac{1}{2}sq - \frac{1}{2}sq^2}{1 - sq}$$

$$\Delta q = q_1 - q = -\frac{\frac{1}{2}sq(1-q)}{1-sq}$$

The strength of the selection will have an effect on the rate at which q is reduced in each generation. As stated before this is only one of the models of selection and the others affect allele frequencies in different ways. Usually selection pressures have a small effect but over several generations this can lead to a loss of an allele unless of course another force such as mutation or migration is acting to recover the allele.

So far we have only considered a single locus situation. Although very few traits are actually controlled by a single locus, they do exist in nature. Mendel's experiments with peas that led to the understanding of what we now know as Mendelian genetics suggested that all visible traits were being effected by single genes. Some important agronomic traits, such as disease tolerance (Cao *et al.*, 2001; Miedaner & Korzun, 2012), are being controlled by single genes. However, many are in fact the effect of a combination of many genes. The actual number of genes controlling a trait is still being debated. Recent studies in wheat revealed that several major QTL with a whole host of smaller interactions are actually affecting the flowering day of year (Buckler 2009). Another contradiction to Mendelian genetics are epistatic effects where the effect of one gene depends on the

presence of one or more 'modifier genes'. Male pattern baldness is an example of an epistatic interaction (Cobb *et al.*, 2010).

The goal of quantitative genetics is to model the variance seen in a trait and then attribute this variance to the underlying genetics. A breeding programme is normally aiming to increase the number of favourable alleles by selecting and using them to create the next generation to raise their frequency in the population, thereby increasing the overall fitness. To achieve this we must know what alleles are affecting any given trait and develop methods to find relationships between alleles and phenotypes. Once they have been discovered, methods for detection within new progeny must be developed.

The interaction between genotype and phenotype is not just of interest to breeding but also to science in general. Understanding the rules that govern such relationships could help us understand the true nature of the genome. Since the genomic revolution has reduced the price of genotyping, more interest has been focused on how we can use new information to reduce the need for phenotyping, which comes with a high cost. Phenotypic selection has long been used in breeding as it was the only resource available to quantify the value of any given genotype. Plants were selected on the basis of performing well to some predefined criteria, such as yield and stress tolerance. Simplistic systems that scored plants on performance are often used to create a range of rankings for a set of genotypes. Even with the simplification of the phenotyping system it is still often very laborious.

On the other hand, recent advances in genome analysis methods such as the use of genotyping-by-sequencing (GBS) allow complex genomes to be represented with a large number of genetic markers (Elshire *et al.,* 2011). The high level of coverage achieved can potentially help to discover how the structure of the genome relates to variations observed

in the phenotypic traits of a given organism. Through careful observation and cleverly constructed experiments we are now starting to unlock links between genotype and phenotype with the hope that breeders can exploit this new information to usher in a new age of genomics led breeding.

Next we will look at two method used to link genotype and phenotype, QTL mapping and genome wide association studies (GWAS).

2.2.2 Quantitative Traits Locus (QTL)

QTL are regions of DNA that have a quantitative association with a phenotypic trait. Breeders and scientists wish to discover these QTL and their effects, and in doing so link genetics and phenotypes. QTL mapping is the method used to do this.

Performing QTL analysis is a lengthy process that requires a large amount of input. Firstly a mapping family must be created. Population size will alter the ability to detect QTL, with higher accuracy being achieved with more progeny (Darvasi *et al.*, 1993). Although one study suggested that above 300 progeny little improvement is seen (Vales *et al.*, 2005). A mapping family will usually consist of a cross between two genotypes that differ in the trait in question. The progeny and parents then undergo genetic analysis leading to the discovery of numerous genetic markers. The higher the marker coverage the greater the chance of finding a marker that is in linkage disequilibrium with a QTL controlling a trait. Once markers have been created, a genetic map must be developed using linkage disequilibrium (LD) to place the markers in the correct order. Once the map is completed, QTL analysis is performed in order to locate the region of the genome that shows high likelihood for a given trait by using observed phenotypic data collected from the mapping family. It is clear that QTL mapping requires great effort in both genotyping

and phenotyping. With data implying only a small number of QTL can be found in each family (Hyne & Kearsey 1995; Kearsey & Farquhar 1998) the cost is high for potentially only a small gain in knowledge. However more recent studies have shown the ability to detect higher number of QTL (Laurie *et al.*, 2004, Buckler *et al*., 2009).

QTL analysis is widely utilised for both scientific discovery and breeding due to its higher degree of accuracy in locating loci compared with other methods currently available. Therefore the methods underlying this approach will now be examined, but before looking at QTL mapping methods, the concept of recombination and linkage disequilibrium (LD) must be examined for they form the basis of QTL analysis. Many methods exist for mapping LD within plant species, see Mackay & Powell (2007) for a review of these methods.

Recombination occurs at meiosis in eukaryotes, and is the process through which crossover between chromosomes leads to novel gametes being created. This cross over acts to change allele frequencies. Recombination can occur at any point in the genome. Recombination frequency (θ) is the frequency with which a single chromosomal crossover will take place between two loci during meiosis.

LD is used to compare the expected and observed frequencies of haplotypes of two loci. From this one can infer how often recombination occurs between these loci, which gives an estimate of how close two loci are in the genome.

Consider two loci, each having two alleles A a and B b, with expected frequency given by

Allele	**Frequency**
A	**p**
a	**q**

B	**r**
b	**s**

where p + q = r + s = 1.

We can potentially have 4 haplotypes whose frequencies are

AB	**pr**
Ab	**ps**
aB	**qr**
ab	**qs**

Assuming that these markers are separated by a defined recombination frequency, θ, we can calculate the coefficient of disequilibrium as

$$D = f_{AB} - pr$$

$$D = f_{Ab} - ps$$

$$D = f_{aB} - qr$$

$$D = f_{ab} - qs$$

where **f** is our expected frequency under HW. The coefficient of disequilibrium is often converted into **D'** and **r²**. To calculate **D'**

$$if\ D<0, \quad D' = D/min(p_A p_B, (1-p_A)(1-p_B))$$

$$if\ D>0, \quad D' = D/min(p_A(1-pB), (1-pA)p_B)$$

To calculate **r²**

$$r^2 = D^2/(pqsr)$$

D' takes a value between -1 and 1, but is often presented as its absolute value. If no recombination occurs between the two loci, **D'** will be 1. **r²** which is sometimes referred to

as Δ^2, ranges from 0 to 1, again 1 implies no recombination has occurred between the two loci.

Genetic mapping relies on the LD to estimate how close two loci are within the genome with the assumption being that if a recombination has not occurred between two loci then they should be closely located in the genome. Using LD, the markers are ordered by the number of observed recombinations. Markers are then clustered together to form linkage groups. This leads to the development of a genetic map, which shows the relative positions of markers to each other.

Interval mapping, is one approach to detect QTL by using pairwise analysis of the markers. For example, consider an F2 population that segregates for two markers with a QTL located somewhere between them. The markers are assumed bi-allelic having values m and M, as is the QTL having values of Q and q. We also assume that the parents were homozygotes for both, the gametes from the F1 have the following probabilities:

$$P_{M_1QM_2} = (1-r_1)(1-r_2)/2$$
$$P_{M_1Qm_2} = (1-r_1)r_2/2$$
$$P_{M_1qM_2} = r_1 r_2/2$$
$$P_{M_1qm_2} = r_1(1-r_2)/2$$
$$P_{m_1QM_2} = r_1(1-r_2)/2$$
$$P_{m_1Qm_2} = r_1 r_2/2$$
$$P_{m_1qM_2} = (1-r_1)r_2/2$$
$$P_{m_1qm_2} = (1-r_1)(1-r_2)/2$$

Where

r_1 is recombination fraction between M_1 and Q, which is unknown

r_2 is recombination fraction between Q and M_2, which again is unknown

From these F1 gametes the probabilities of the F2 genotypes can be defined. For example

$$P_{m_1 m_1 QQ M_2 M_2} = [(1-r_1)(1-r_2)/2]^2$$

A likelihood value can then be derived from these, and through maximising, an estimate of the effect for the QTL can be calculated. This can also give us an estimate of the recombinant fraction r_1. Due to the prior creation of a genetic map, we already know the recombination fraction between M_1 and M_2 so r_2 can be calculated using the estimate of r_1. This approach of maximum likelihood (ML) is adopted in many software packages. They calculate the maximum likelihood at regular intervals between markers. Maximum likelihood scores can be plotted against the location in the genetic map and used to find where QTLs potentially lie. Maximum likelihood is simple to apply but in a situation with large marker datasets can be computationally intensive.

Haley and Knott (1992) created an interval mapping method that uses least squares regression. The results are identical to ML estimations but can be solved with shorter computation time. It works by calculating the additive (a) and dominance (d) effects for the QTL. We assume the mean values of each QTL (μ) with two alleles q and Q are

$$\mu_{QQ} = \mu + a$$
$$\mu_{Qq} = \mu + d$$
$$\mu_{qq} = \mu - a$$

And we have two markers M_1 and M_2 flanking the QTL, each marker having two alleles m and M. The regression then takes the form

$$z_i = \mu + ax_1 + dx_2 + e_i$$

Where z is the observed phenotype, and x_1 and x_2 are the probabilities of the QTL, x_1 for the homozygotes and x_2 for the heterozygote. We can then take the mean for any genotype, for example $M_1M_1M_2M_2$

$$\mu_{M_1M_1M_2M_2}=\mu+ax_1+dx_2$$

We then need to calculate the mean for this marker pair, and equate the x_1 term to the QQ and qq classes and the x_2 term to the heterozygote Qq. If we again look at the conditional probabilities of the QTL in an F2 population for the class $M_1M_1M_2M_2$

$$P_{M_1M_1QQM_2M_2}=[(1-r_1)(1-r_2)/2]^2$$
$$P_{M_1M_1QqM_2M_2}=2[(1-r_1)(1-r_2)/2][r_1r_2/2]$$
$$P_{M_1M_1qqM_2M_2}=[r_1r_1/2]^2$$

We next sum the probabilities for all possible values giving us

$$P_{M_1M_1M_2M_2}=\left(\frac{1-r_{M_1M_2}}{2}\right)^2$$

where

$$r_{M_1M_2}=recombination\ fraction\ between\ M_1 \text{ and } M_2$$

And then, using the definition of conditional probability, we can calculate the probability of each QTL class given the probability of the marker genotype classes

$$P_{QQ|M_1M_1M_2M_2}=(1-r_1)^2(1-r_2)^2/(1-r_{M_1M_2})^2$$
$$P_{Qq|M_1M_1M_2M_2}=2r_1r_2(1-r_1)(1-r_2)/(1-r_{M_1M_2})^2$$
$$P_{qq|M_1M_1M_2M_2}=(r_1r_2)^2/(1-r_{M_1M_2})^2$$

Now returning to the regression of our marker pairs, we can use these probabilities to work out the expected mean.

$$\mu_{M_1M_1M_2M_2}=\mu+\alpha\left[\frac{(1-r_1)^2(1-r_2)^2-r_1^2r_2^2}{1-r_{M_1M_2}}\right]+d\left[\frac{2r_1r_2(1-r_1)(1-r_2)}{(1-r_{m_1m_2})^2}\right]$$

This process is repeated for each possible marker pairs, leaving a set of coefficients to be fitted. As with ML, the interval (r_1) is varied and the result with the lowest error in that range is the best QTL placement. QTL maps are usually plotted as a LOD score along a linkage group. LOD stands for a logarithm of odds and compares the likelihood of an event

occurring compared to the likelihood of it occurring by chance. For QTL mapping using regression an equivalent of a LOD score is calculated from $SS_{regression}/SS_{total}$ which can then be plotted against the location in the genetic map, same as the ML method.

It is clear that QTL mapping requires large amounts of computation in order to best locate a potential QTL. All of this is based upon linkage, but linkage is not always simple to detect. Markers which are located far apart on a chromosome or even on different chromosome can appear to be in linkage. This can lead to a false QTL being detected. There is also issues such as ghost QTL where two closely located QTLs might appear as one single large QTL, whereas there are actually two smaller QTL affecting the underlying trait.

2.2.3 Genome Wide Association Mapping (GWAS)

Genome wide association studies (GWAS) aim to utilise whole genome genotyping methods such as GBS in order to find associations between phenotypes and markers. Associations are created by looking for variance in the genetics which matches variance seen in the phenotype. Large numbers of markers (usually SNPs) are needed for GWAS studies. By using large numbers of markers there is a greater chance of finding markers in LD with the QTL.

Tests for association are performed on each SNP. For quantitative data generalised linear models are used, most commonly ANOVA. With a null hypothesis being that the SNP has no effect on a given trait. When thousands of SNPs are analysed in GWAS there are thousands of tests that must be performed. This multiple testing increases the chance of false positives being detected just by chance. Commonly the p-value is set to 0.05, a 5% false positive rate, in a normal statistical test. However over the thousands of tests

performed in GWAS the cumulative likelihood of finding a false positive increases. Therefore a more strict significance threshold must be selected. One of the simplest approaches is to apply Bonferroni correction calculated as the current false positive rate over the number of tests. So for our 0.05 rate, if we had 500,000 SNPs, our new threshold would be 10^{-7}. Alternative methods for setting a threshold include include permutation tests.

GWAS analysis assumes that it is the common variants that will explain a significant amount of the variation seen rather than the rare alleles. Any SNP discovered to be in association with a trait is assumed to be located near to a gene controlling that trait. It uses the theory that each locus will in fact only account for a small amount of variance for a given trait, so will not attribute large amounts of variance to a given marker. As with all genomic studies, GWAS need a large number of genotypes within the population in order to detect the traits. It should also be noted that any marker found to be significant is only an association and does not imply causation.

There exist several programs to perform GWAS analysis which allow for the large number of statistical tests needed and can perform calculations to discover the significance threshold. EMMAX is one such program (Kang *et al*., 2010).

Kinship plays a large role in GWAS studies, as stratification of samples can exist. Population stratification is the presence of common allele frequencies within subpopulations caused by a common ancestry. For example there might be genomic differences in population due to geography, such as between two countries, where inbreeding will be mainly contained within the country itself with only a small proportion of migrants passing alleles between the populations. Hence the kinship of all genotypes within the study needs to be calculated first. Tools such as EMMAX can make use of these

kinships in order not to detect structural genome changes that related to these sub populations, so that genes actually relating to the trait across these populations can be found (Turner *et al.,* 2011; Ma *et al.,* 2012a)

For the last decade, GWAS has been extensively used in human biology when looking for genetic markers that show association with diseases or health related conditions (Scott *et al.,* 2007; Welcome Trust, 2007). GWAS has also been applied to many crop species, such as rice, barley, wheat, maize, Lolium and *Miscanthus* (Skøt *et al.,* 2005; Cockram *et al.,* 2010; Neumann *et al.,* 2010; Kump 2011; Tian *et al.,* 2011; Zhao *et al.,* 2011; Slavov *et al.,* 2014).

Although GWAS can reveal markers related to many traits, it is still unable to account for all the variance, leading to the term missing heritability which refers to the variance that has been shown to have a genetic component but is still not totally accounted for in GWAS studies (Manolio, 2009; Brachi *et al.,* 2011). GWAS, as previously mentioned, only considers common loci that have a small effect on any given trait. There are two types of loci that are not considered – rare alleles with small effect, and more commonly occurring alleles but with a large effect. Large effects of low frequency are more often discovered in studies using QTL analysis but the rare low effect ones are not often seen in any type of genetic study. This is due to that even with large sample sizes seen in GWAS studies there may still be alleles unaccounted for due to low frequency.

Another main problem for all types of genomic analysis is the absence of high quality full genome in many species. However with the coming age of high throughput technology it is possible to generate large marker datasets that provide good coverage of the genome, allowing for a greater chance of finding a marker which is in LD with the trait in question. This could potentially increase the likelihood of being able to detect rare

alleles by including more genotypes into the analysis without the cost increasing exponentially.

2.2.4 Genomic Selection (GS)

This final section will look at a recent theory that aims to make use of whole genome association data to inform breeding decisions. Genomic selection (GS) is a marker assisted selection (MAS) approach for breeding, in which genetic markers which cover the whole genome are used so that all QTLs should be in linkage disequilibrium with at least one marker. The theory is to assign a breeding score to the genome as a whole, allowing each genotype to be quantified for its potential breeding value. This value is then used to estimate the performance of new progeny from their genetics without the need for phenotyping. Let us first look at the standard breeding cycle (Figure 2.2).

It is seen in the standard breeding cycle that all decisions hinge around field evaluations for all genotypes. Genotypes from crossing or germplasm collections undergo evaluations in the field. Phenotyping in field is an expensive and time consuming process requiring skilled staff. In some perennial crop species where a plant needs to reach maturity before it can be evaluated, which may take several years, this presents a clear bottleneck.

Genomic selection aims to reduce the time taken for evaluation. This is done by creating prediction models that predict the phenotype or breeding score of an individual from only its genetics. The theory being that this will minimise the time taken for each round of the breeding cycle (Figure 2.3).

In the genomic selection breeding cycle selection is now performed using a predictive model, meaning that crossing and parental selection can now be done without the need for

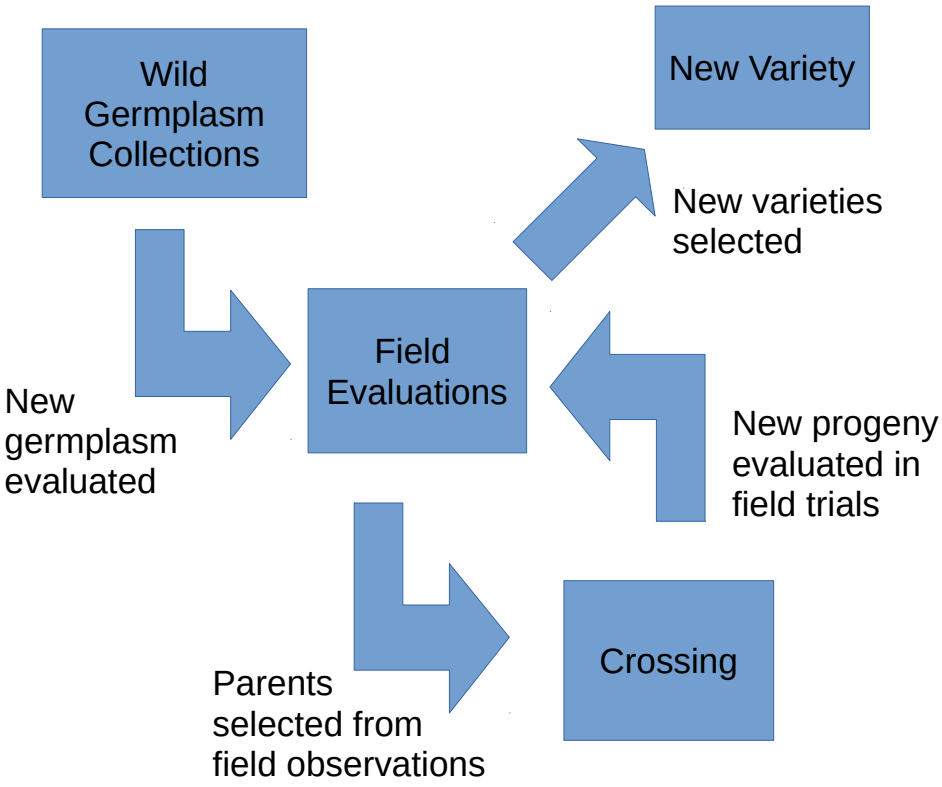

Figure 2.2: A standard breeding cycle where new crosses and wild germplasm are evaluated in field trails by phenotypic observations. Selections are then made by the breeder as to which genotypes should be used in crossing or released as varieties. It is clear that all decision hinge upon the field evaluations. In a perennial crop with a several year life cycle, the time taken to get evaluations means a bottleneck in the breeding programme.

field evaluations. However, new progeny do need to be genotyped, but this is usually a cheaper option to phenotyping and much faster. Field evaluations are still needed for new varieties, and also to provide new information for the model, such that it can be retrained as performance decreases. The decrease in performance is caused by the genetics of the underlying population changing as breeding processes, this can act to alter allele frequencies in a way that the model has not previously seen. This may effect the accuracy of the model and therefore require a new model to be trained utilising both the historical and new data. It should be clear that a breeding program should progress faster using GS as long as a good model can be developed.

The question is how we build a model that will provide good predictions of our new

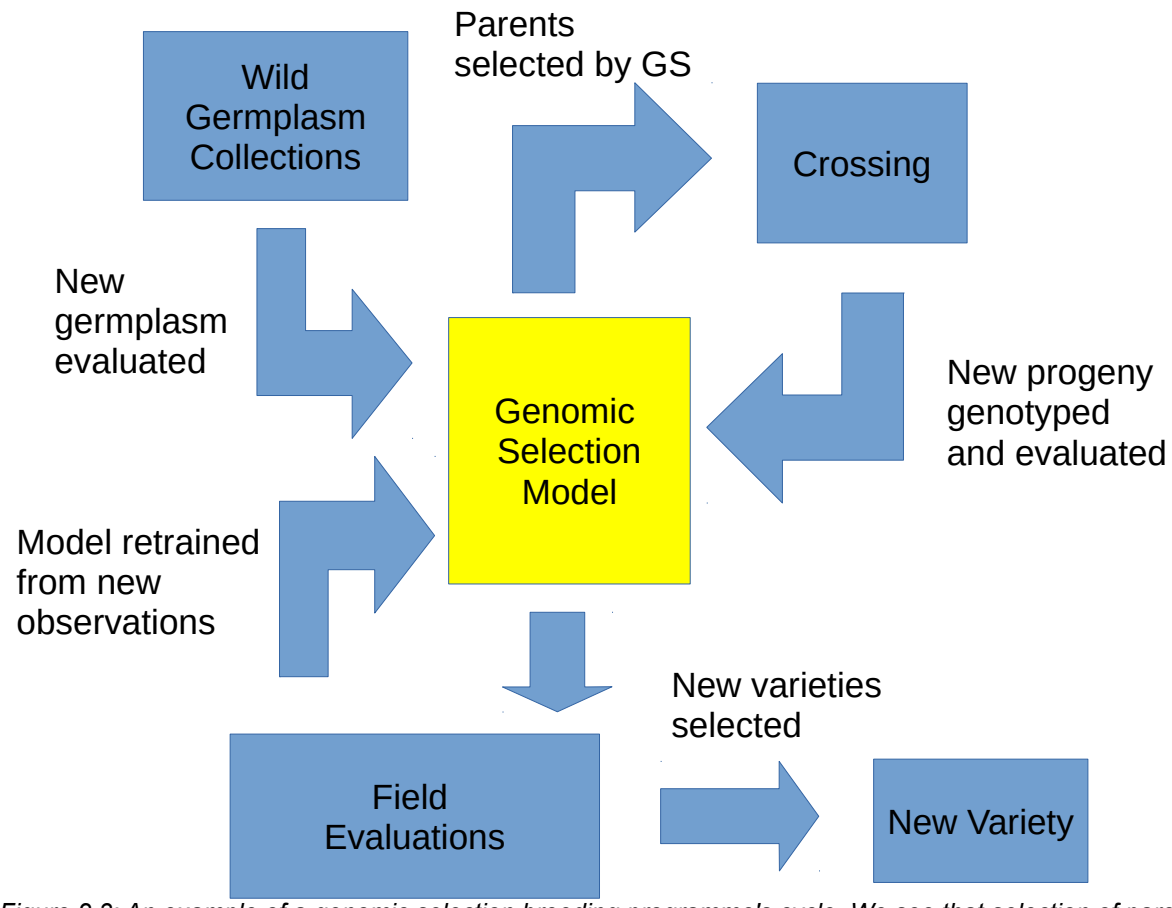

Figure 2.3: An example of a genomic selection breeding programme's cycle. We see that selection of parents for new crosses no longer requires field evaluations. Although new progeny are now require genotyping, this is often more economical than phenotyping. All selections are now performed using the genomic selection model. Field evaluations are still required but they are now only used to retrain the model and evaluate new varieties. This means that the number of field evaluations is decreased and no longer causing a bottleneck in crossing and progeny evaluation.

progeny. GS was first proposed by Meuwissen *et al.* (2001) and has been extensively used in cattle breeding (Luan *et al.,* 2009; Hayes *et al.,* 2009). It has more recently been used in crop breeding programmes (Heffner *et al.,* 2009; Heffner *et al.,* 2010; Sorrells *et al.,* 2011; Hayes *et al.,* 2013).

Many approaches exist for working out the breeding value of individuals. Meuwissen *et al.* (2001) presented four methods in their paper: least squares regression (LS), BLUP, BayesA and BayesB. Meuwissen created a simulation of 1000cM genome with an effective population size of 100 to make sure that linkage existed between markers and QTL. The

results from the study showed that LS performed the worst and also was unable to detect the simulated QTL. The BLUP method performed better but it was pointed out by the author that it incorrectly assumes equal variance for all loci. BayesA and BayesB performed the best with the latter being the best with approximately 16% better performance than that of the BLUP method. The BayesB method also performed better when the number of phenotypes was lowered with only a small amount of loss in accuracy whereas LS and BLUP performance was negatively affected by the lower numbers of phenotypes. In another study Luan *et al* proposed that the size of the training set will influence prediction accuracy. This, as suggested, may be an effect caused by the heritability of the trait within the population (Luan *et al.,* 2009).

Since Meuwissen's first study other approaches have been applied to genomic selection – Heslot *et al.* (2012), for example, tested several approaches on a variety of data sets. This study also looked at the performance of several machine learning methods for predicting breeding values. The methods tested included random forest, elastic net, a variety of Bayesian approaches and support vector machines.

Jannink *et al*. (2010) investigated how genomic selection can be used in practical breeding programme by drawing comparisons between several methods. They addressed the weaknesses of MAS, pointing out that it is costly to generate mapping populations for MAS, which leads to small sample size and therefore underpowered analysis. Small effect QTLs are often missed and the use of bi-parental populations limits the variation observed when compared to the breeding population. They also point out that association mapping also has disadvantage of biased effect estimations leading to poor performance in prediction. As the cost of genotyping decreases and the cost of phenotyping remains higher than the cost for genotyping, GS looks more appealing. Jannink *et al.* also looked at

several methods that can be used to perform GS. Non-linear methods have the potential to detect epistatic and non-additive effects. Based on many simulations the authors theorized that if the objective is to accelerate a breeding cycle, genomic selection provides a better solution. This means that GS could be useful in perennial crops where breeding cycles tend to be long. The authors also state that in bi-parental populations GS can outperform phenotypic selection even with small numbers of individuals. In fact GS would actually require a lower number of markers than conventional methods used in bi-parental studies, although separate modules would be needed per family. A lot of the studies used replicated genotypes but they argue that several studies highlight that the accuracy of GS could be maximised by evaluating un-replicated genotypes. If GS was to be applied it would mean intensive phenotyping is needed to train prediction models, and training populations should, maximize marker variance. It was also suggested in the study that models need to be retrained as populations develop.

Having the ability to account for the smaller effect QTL, GS is a powerful methods when compared to MAS or association mapping. However, it is not clear which GS methods in particular are best. Between BayesB and Ridge Regression the accuracies of results may vary according to family design. The number of markers needed is also another area in need of more in depth study. Also the best design for trials is still under some debate, but regardless of these questions, GS provides another powerful method for crop improvement.

2.3 Computational Crop Modelling in Plant Breeding

2.3.1 Crop modelling: theory and practice

Crop modelling is used for both prediction and understanding of interactions between

many crop features such as genetics, phenotypical traits and physiological traits (Hammer *et al.,* 2002). Physiological traits related to characteristics that determine how a plant functions such as the ability to conserve water, whereas phenotypical traits look at visible measurements such as plant height or flowering stage.

Traditionally, crop modelling has been utilised in many species to provide predictions for growers in order to select the best variety for a particular environment. MISCANFOR (Hasting *et al*. 2009a) is one of such model established for *Miscanthus.* It provides yield predictions for the naturally occurring sterile hybrid *M. x giganteus* based upon observations taken from various climate and soil conditions.

Models are also used to estimate traits that are difficult to measure directly. In *Miscanthus* modelling has been used to estimate physiological traits such as radiation use efficency (RUE) (Davey *et al., In Preparation*). By measuring traits such as leaf area index (LAI) and plant dry weight and using mathematical models the authors were able to calculate the RUE of various *Miscanthus* species. Due to the intensity of the measurements needed to perform this type of study they are usually limited to a small number of genotypes.

The problem with these types of models is that they require a lot of phenotypical or physiological measurements that are both costly and time consuming. Alternatively, genomic modelling in crops requires large numbers of genotypes to detect the variations that exist across multiple genomes that cause differences in the phenotype. Methods such as QTL mapping, GWAS and GS develop models that examine the genetic information in order to link them to phenotypical or physiological measurements. Although most of these studies focus on phenotypical links (Huang *et al*., 2012; Pasam *et al*., 2012; Morris *et al*., 2013; Li *et al*., 2013). It is difficult to get large amounts of physiological measurements

needed to develop the models for linking genetics to physiology.

As mentioned previously, one of the most important traits of interest for energy and food crops is yield (Hasting *et al.,* 2009a; Robson *et al.,* 2013; Jensen *et al.,* 2013). Linking yield with genetics has had mixed results consisting of many small and large effect QTL with some involving epistatic interactions (Kumar *et al*., 2006; Xing and Zhang, 2009). This could be due to yield being a complex trait affected by a combination of both phenotypical and physiological traits (Marcellis *et al*., 1998).

2.3.2 Modelling methodologies

Modelling plant growth and genetic and environmental effects has long been established in crop development and prediction as tools to inform breeders and growers on the best course of action. The following sections will outline some commonly used methods when modelling crops. These do not link genotype and phenotype directly as the methods discussed earlier but are still important for understanding how plants are affected by the environment and for attributing variances to genetic factors.

Process Modelling

Process models mainly focus on physiological studies where variables may be hard or impossible to measure. They are utilised to explain and model the physiological processes within a plant. Mathematical equations are used and missing terms are calculated from observed data. Many models have been developed to understand how crops grow and interact with their surroundings. LAI is an example of a measure used to estimate the physiological behaviour of a plant. LAI, leaf area index (Watson 1947), uses an assumed mathematical model to estimate the area of leaf a plant uses to capture the incoming light. LAI is a dimensionless measure of leaf area per ground surface area. If LAI

is 0 there is no leaf cover and the index increases as more leaves develops to capture more light throughout the growing season. LAI can be measured directly by measuring width and length of each leaf to calculate the area of light interception (Breda, 2003). This measure is used extensively in plant modelling and is also needed for Penman-Monteith (Allen *et al.,* 1998) equation for net evapotranspiration and calculations for radiation use efficiency.

Many mathematical process-based models like this can be linked together and parameterised from the observations taken in the field. These models can then be processed via time steps that can be used to model how a plant responds to varying inputs. This can then be used to estimate behaviours from the observations by adjusting some of the parameters. Another use of process models is to estimate the effects of changing a system that might not be possible in the real world. Parameters can be changed to reflect stresses that might be otherwise hard to produce in order to predict the response from the plant.

Phenotypical Variances

The phenotype of an organism is considered to be a function of two factors, genetics and environment, and observed phenotypic variation is also a function of these. This can be described as follows

$$V_P = V_G + V_E$$

Where

V_P is the phenotypical variance

V_G is the genetic variance

$\mathbf{V_E}$ is the environment variance

Based on this formula, mixed models can be used to break down phenotypic variation into its environmental and its genetic components. In order to work out how much each is affecting the variance a trial must be carried out with genetic clones and observations must be taken on all. This data can then be used to build a model which will tell how much of the variance is explained by genetics and how much is explained by environment.

Through the application of randomised trial designs, environment effects can be accounted for using row, column and also block effects. This can help eliminate problems caused by uneven nutrient distribution or changes in field topology which could potentially effect the traits being measured. This allows us to understand how much of a trait is due to genetics. Using this data we can then get the broad sense heritability ($\mathbf{H^2}$) of a given trait which is calculated as

$$H^2 = \frac{V_G}{V_P}$$

This is a score of one to zero that explains how much of a trait is genetic and how much is environmental in the given environment from which the observations were taken. For *Miscanthus* broad sense heritability has been shown to range from 0.89 for flowering to 0.48 for plant stature (Slavov *et al*. 2014). Genetic variance can also be further broken down into several factors. If we assume there are three genetic factors at play, the additive effect of a gene (the difference seen in the phenotype from that gene), the dominance effect of the interaction of the alleles and epistatic interactions between genes within the genome. The formula above can then be expanded to

$$V_P = V_A + V_D + V_I + V_E$$

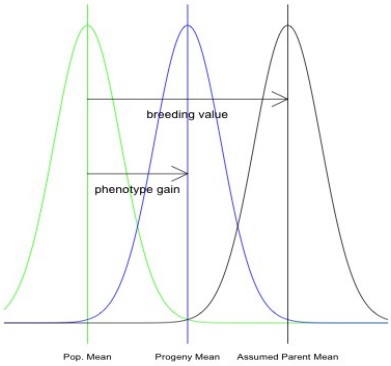

Figure 2.4: Example distributions for breeding populations illustrating how breeding values are calculated. In green we see the distribution of the breeding population, blue is the distribution of the progeny, and black shows the expected distribution of the parent for whom the breeding value is to be calculated for.

The parameters $\mathbf{V_A}$, $\mathbf{V_D}$ and $\mathbf{V_I}$ are representing additive, dominance and interaction effects respectively.

Another measure of heritability is narrow-sense heritability, this only accounts for variance from additive effects

$$h^2 = \frac{V_a}{V_p}$$

Calculating the additive variance is difficult with current technology therefore instead an estimation is achieved by looking at the relatedness between individuals. One way to estimate narrow-sense heritability is to perform a regression between the offsprings phenotypic value and the parents phenotypic value. The narrow-sense heritability is then the slope of the regression ($\mathbf{b_{OP}}$).

There are multiple ways to perform this regression, first is using single parent offspring regression, where O is the mean of the offspring and P is the mean of the parent.

$$b_{OP} = \frac{V_A}{2V_P} = \frac{1}{2}h^2$$

Alternatively regression can be performed between the offspring mean phenotype and the midparent mean phenotype. If we assume O is the mean phenotypic value of the offspring, and P_1 and P_2 are the means of the parents. If $\overline{P}$ is the mid parent value where $\overline{P} = ½ * (P_1 + P_2)$ then narrow-sense heritability is

$$b_{O\bar{P}} = \frac{V_A}{V_P} = h^2$$

2.3.3 Modelling as decision-support tool in crop breeding

The use of modelling in crop breeding has long been established. The first example of this is phenotypic selection, where models were built using phenotype data to predict the phenotype of new progeny based on the prior knowledge of several previous generations.

Breeding scores, or the potential net gain from using a particular genotype in a breeding program, is one of the most frequently used methods in crop breeding. This score can be calculated via genomics, such as in genomic selection, or by phenotypic observation. We discuss the phenotypic approach in what follows. Calculation is performed on a single phenotypic trait and is based on comparing improvement in the phenotype compared to the base population. In this example we assume a base population of 50 genotypes and we want to calculate the breeding value of one individual. Firstly the phenotypic trait in question, which is assumed to be normally distributed with mean zero, is measured for all the individuals. Once calculated, the phenotypes will be normalised to a mean of zero.

Next using the genotype which we wish to calculate the breeding score for, we create crosses with as many of the other genotypes as possible and observe the progeny. We then normalise the base population mean to zero and the progeny mean by the population

mean. Then we calculate the deviation of the progeny from the population. Let us assume that the new population has a mean of 1.5 after normalisation. The breeding value is calculated as twice the deviation of the progeny mean, so for the example the breeding value of the genotype is 3 (Figure 2.4). This is based on the assumptions of additive genetics, so the progeny would be half way between the mean of the population phenotype and the mean phenotype of the genotype. The variation around the mean is attributed to environmental interactions.

We can attribute the change in breeding values to loci. Firstly the mean of the population must be calculated according to gene frequencies. If the mean of our population is 0 for a given phenotype, we can attribute values to the genotypes observed. Assume the A_1 allele confers an increase in the phenotype, then

Genotype	Assigned Value
A_2A_2	-a
A_1A_2	d
A_1A_1	+a

According to the definitions of dominance if A_1 displayed complete dominance then d = +a, and if A_1 displayed no dominance d = 0 (Figure 2.1). Finally if there was overdominance in the heterozygote d > +a. Overdominance is where the phenotype of the heterozygote is greater than either of the homozygote. By calculating the gene frequencies as done previously, the values can be obtained as

Genotype	**Frequency**	**Assigned Value**	**Freq * Value**
A_2A_2	q	-a	$-q^2a$
A_1A_2	2qp	d	2pqd
A_1A_1	p	+a	p^2a

Hence we calculate population mean to be:

$$Pop.Mean = a(p-q) + 2dpq$$

Since we are assuming an additive model, if we were to account for multiple allele the population mean becomes the sum of all the allele contributions. For each gamete we can calculate the mean value for the genotypes produced, for our single loci example this would be

Gamete	Mean Value
A_1	pa + qd
A_2	-qa + pd

From this we can deduce the population mean and the effect of A_1 can be represented as α_1

$$\alpha_1 = pa + qd - [a(p-q) + 2dpq] = q[a + d(q-p)]$$

and for A_2

$$\alpha_2 = -p[a + d(q-p)]$$

The breeding values, defined as twice the contribution to the mean of respective allele, can now be calculated to be:

Genotype	Breeding Value
A_2A_2	$2\alpha_2$
A_1A_2	$\alpha_1 + \alpha_2$
A_1A_1	$2\alpha_1$

As mentioned before, these models are based on the assumption that all alleles are additive and as such do not account for allele interactions, commonly referred to as epistatic effects. Using single or multiple loci and summing these effects would give the

new mean. However, this implies that we know the additive effects of each locus. In reality knowing all the loci affecting a trait is difficult and knowing the effect of each locus has on the trait would be impossible with current technology. Instead we use estimations of the additive effects such as narrow-sense heritability.

Narrow-sense heritability can be used to estimate the response to selection when breeding for a particular trait. This is expressed as the breeders equation

$$R = h^2 S$$

Where

R is the response to selection

h^2 is the narrow-sense heritability of a trait

S is the selection differential (The mean phenotypic value of the parents, expressed as deviations from the population mean).

The breeder's equation can be considered for marker-assisted selection (MAS) as

$$R = ih\sigma_g$$

where σ_g is the genetic variance, h is the narrow-sense heritability of the trait and i is the standardised selection differential.

Of course the goal of using MAS for breeding is to outperform the phenotypic selection; otherwise there is no advantage to make the extra effort of QTL mapping. Let us consider the two forms of selection

$$\begin{array}{ll} \textit{Phenotype only} & R = ih_p^2\sigma_p \\ \textit{Markers' only} & R = ir_g h_m h_p \sigma_p \end{array}$$

where the subscript p refers to the phenotype heritability and variances and

subscript m is the same for the marker, and finally r_g is correlation between the index of a score based on the markers and the phenotype. We can now compare marker-assisted selection to phenotypic selection. For MAS to be the better method we need:

$$ir_g h_m h_p \sigma_p \quad > \quad ih_p^2 \sigma_p$$

which can then be simplified to

$$r_g h_m \quad > \quad h_p$$

And since $h_m^2 = 1$ (assuming no errors in genotyping)

$$r_g \quad > \quad h_p$$

$$r_g^2 \quad > \quad h_p^2$$

So for MAS to outperform phenotypic selection, the squared correlation coefficient between marker index and genotype must be higher than the heritability of the given phenotype. MAS would therefore be unbeatable if all QTL of a trait were known and these effects were quantified without error and were correctly weighted in the index, as r_g^2 would be 1. However, in practice this is not possible to achieve in QTL studies where majority of them are accounting only for a subset of the total variance due to many small effect unknown QTL.

Genomic selection is another recent theory in using models for breeding. Its goal is to create a genomic model using whole genome genotyping methods to create a representation of the genotype that can be used to model the effects of multiple loci on a trait. The idea being that the model will aid in the decision of which parents to use, which progeny to select for testing and reduce the amount of effort required for phenotyping and evaluation of crops.

2.4 Genotyping and Phenotyping Methods

2.4.1 High-throughput methods and next generation sequencing

In order to be able to understand the nature of gene-trait associations, both genotype and phenotype must be accurately captured and quantified. Large numbers of genotypes must be observed in order to capture the necessary data to facilitate the implementation of the methods discussed earlier.

The advancement of next generation sequencing (NGS), generating gigabytes of data, has revolutionised the world of genetics providing us with huge genomic resources for genetic research (Metzker *et al*., 2010). RNA and DNA data can now be processed giving us insights into the underlying nature of the genome and transcriptome of any organism we wish to study. Genotyping-by-sequencing (GBS) is one of the methods to exploit this NGS technology, through which an organism can be genotyped, creating thousands of single nucleotide polymorphisms (SNP) markers that can be used in mathematical models for molecular assisted breeding as previously discussed. The use of GBS method for complex organisms such as higher plants is presented in Elshire *et al*'s paper (Elshire *et al.*, 2011) and a brief description of this process is outlined below.

Firstly the DNA samples must be extracted from the organism. This DNA is then fragmented using a restriction enzyme. Different enzymes used in the DNA digestion will lead to discovery of different SNPs due to enzyme cutting the DNA at different loci. Barcode sequences are then attached to the DNA samples so that they can be identified later. Afterward, the samples are sequenced using sequencing technology such as Genome Analyzer II (Illumina, Inc., San Diego, CA). The resulting sequences are then aligned to the same genotype using the barcoding DNA attached earlier. The results are

filtered to make sure that sequences are found in multiple reads to remove any false reads. SNPs will then be mapped using a threshold of $p<0.0001$ for the binomial test.

Even with all of these new high throughput methods in genotyping there is still a bottleneck in data acquisition, phenotyping. The process of phenotyping is time consuming and requires skilled workers to be able to observe and collect data. However, image analysis techniques, such as those used in the national plant phenomics center at IBERS, have the potential to provide a high throughput method for phenotyping. The system consists of a conveyor belt that moves plants through a series of imaging chambers, collecting large volumes of data. This process can handle thousands of plants within a short period of time but these plants must be relatively small in order to fit on the conveyor system and be pot grown. However, this approach does not consider the environment effect the plant will experience in the field. Methods for high throughput phenotyping in the field are still in their infancy.

2.4.2 Genotyping Miscanthus

Genotyping of *Miscanthus* has been performed using the genetics method discussed in Elshire's paper (Elshire *et al*., 2011). Two populations have been genotyped using this method at IBERS and formed the main datasets which were analysed in the research described in this thesis.

The Mx2 mapping family has been designed for genetic mapping and QTL analysis on *Miscanthus* flowering time. This mapping population consisted of 185 genotyped progeny of two flowering *M. sinensis* genotypes, one with early and another with late flowering times. Approximately 17,000 SNP markers were identified from which 3785 markers were then selected that could be confidently rated for the identification of the

alleles in the population (Ma *et al*., 2012b). The genetic map was then created using this dataset as outlined in Ma's paper. The analysis revealed that *Miscanthus* has high synteny with *Sorghum*. Therefore *Sorghum* was used to assist in the construction of the genetic map.

A second population was created to study the diversity within the wild *Miscanthus* germplasm collection at IBERS, from this population 244 accessions were selected for analysis. To date 179 of the 244 accessions have had markers detected using GBS technique (Elshire *et al.,* 2011). In this population 3777 SNP markers were discovered. The resulting SNP markers have then been mapped on to the *Sorghum* physical map and genome. This allowed for comparison between markers in the *Miscanthus* and *Sorghum* populations.

The high synteny between *Miscanthus* and *Sorghum* provided an additional potential data source and comparisons between *Sorghum* and *Miscanthus* could be performed. Most of the markers generated in either of the GBS studies could be fully mapped onto the *Sorghum* physical genome (Patterson *et al*., 2009), providing access to known *Sorghum* QTL, potential gene models and a whole range of homologs that have been found between *Sorghum* and other crop species. They could be used to confirm the accuracy of the machine learning approaches when detecting relationships between genotype and phenotype. This was done by searching the literature discussing effects of genes or their homologs on the traits under consideration. *Miscanthus* is an undomesticated crop with very little information about its genetics, whereas *Sorghum* and other closely related crop species such as rice, maize and B*rachypodium*, have been studied in great detail and so can be utilised in order to provide validation of the associations.

2.4.3 Phenotyping Miscanthus

In what follows we outline the protocols used for the phenotypic observations taken for the analysis presented in this thesis.

Flowering Time

Phenotypic data on flowering time was one of the most extensively used for this research project. Jensen *et al* investigated flowering time in *Miscanthus* species (Jensen *et al.,* 2011a) in which flowering time was recorded on a scoring system based upon a visual observation of the stage in the flowering cycle system. Two different scoring systems were used when monitoring *Miscanthus* flowering. The first system, described in the paper of Jensen et al (Jensen *et al*., 2011a), is shown in Table 2.2. The second scoring system is a five values scoring system described in Table 2.3. The reason two systems were used was to measure two different aspects of flowering, with the first system measuring flowering intensity and the second looking for a time series for the various stages needed to complete flowering. The first two scores are the same in both systems, so when comparing across multiple datasets only flag lead and panicle emergence were used.

Score	Definition
1	Flag Leaf emergence
2	Panicle Emergence (2cm)
3	50% stem flowering
4	80% stem flowering

Table 2.2: Flowering scoring system used in several flowering trials at Aberystwyth

Score	Definition
1	Flag Leaf emergence
2	Panicle emergence (2cm)
3	Anthesis
4	Seed Set
5	Flowering Completed

Table 2.3: Flowering score used by the Miscanthus Breeding program at IBERS

Canopy Height

Canopy height is another important trait that contributes to the yield of Miscanthus (Gonza *et al.*, 2001; Robson *et al.*, 2013). It is measured from the ground to the top of the bulk of the canopy and for *M. sinensis* this is where the leaves level out. As for the *M. sacchariflorus*, due to the upright leaves, it is normally taken to the middle of the tallest leaf to allow a fair estimation as to where the bulk of the canopy ends. Tallest stem is measured to the highest ligule leaf.

Stem Diameter

The width of a stem in plant is measured as stem diameter. To do this several measurements are taken and an average is calculated. Stems are measured at approximately half way up the canopy. Noduled stems are taken approximately half way between two nodules, as this is more representative of the stem as a whole.

Base Diameter

Base diameter is measured between the two farthest apart stems, this is done at ground level. However, this measurement protocol cannot be used for some *M. sacchariflorus* plants due to their creeping nature which means that sometimes it is difficult to distinguish stems from two neighbouring plants.

Stem Count

For smaller plants this is taken as an actual count of the number of stems that can be attributed to a given plant. In larger plants (those that clearly contain several hundred stems, and where time was a limitation during remote site visits, for example) the measurement takes the form of an approximation estimate. When an estimate is used the plant is split by eye into quarters and then stems in one quarter are counted and multiplied by 4 to estimate stem count. In some mature *M. sacchariflorus* it sometime can be difficult to distinguish which plant a stem belongs to, as some rhizome can grow outwards horizontally from the planting location. Where several *M. sacchariflorus* plants are closely located, this can mean stems will fill the space between the plants making the count more difficult. In these instances a stem is included in the count only if it can be confidently assigned to a plant.

Age

Being a perennial crop *Miscanthus* has a maturity requirement, whereby it may take several years for a mature phenotype to be displayed. Therefore the age of the plant may be useful when comparing phenotypes. With this in mind a system was created in the database which requires the date of planting to be recorded. This allows one to infer a plant's age at any date of interest. This is always considered as number of years since planting.

Moisture Content

To calculate the moisture content of a plant a sub sample of the harvested plant material is weighed. The sample is then put into a drying oven for 24 hours and is the reweighed. The loss in weight is calculated as a percentage giving the moisture content of the plant.

2.5 Data Collection and Handling for Miscanthus Flowering Time and Growth

2.5.1 Software Usage and Development

The datasets used in this research were generated from a variety of sources with various file formats in which measurements were recorded using different protocols. Before any analysis could be performed, all data were collated and standardised to facilitate the machine learning and statistical analysis. Analysis performed in this research was done using open-source software, mainly the statistical language R (R Core Team 2013), and the machine learning suite WEKA (Hall *et al*., 2009). R is mainly a command line interface tool consisting of a series of libraries, known as packages, through which difference functions can be implemented. A list of R packages used is provided in Table 2.4. Weka is a Graphical User Interface (GUI) interface developed using Java, through which many of the machine learning algorithms and statistical methods can be applied.

Package Name	**Usage**
lattice	Graphics suite for plotting
randomForest	Implementation of Breiman's random forest algorithm
SPRINT	Parallelisation of several common R functions, and also random forest
Rweka	Interface for using WEKA functions directly in R
ggplot2	A graphics library sometime utilised instead of lattice

Table 2.4: List of R packages used

Python (Rossum, 1995), a scripting language, was employed for data handling and management in this research. A SQLite database was created to store phenotype, genotype and environmental data. The use of a single database meant that data could be standardised during import so analysis could be performed across data sets. Building a

simple object to database mapping framework allowed for the system to be flexible so adding new phenotypes could be easily accommodated into the database without much modification (Figure 8.5). This flexible framework allowed for new experiments and data sets to be added to the database with ease in order to either link data together (such as phenotypical and genomic data) or to perform analysis on new datasets. This database was only utilised by myself, the *Miscanthus* breeding programme has its own database know internally as MSCAN. MSCAN has been developed by an in house software developer to contain all the information pertaining to breeding of *Miscanthus* at IBERS. It provides data storage for information such as phenotypical measurements, while also keeping track of crossing operations and stock locations of each *Miscanthus* accession. I developed the separate Python based database which could take data from the database to allow me the flexibility to perform complex analysis without having to alter the structure of the group database. Also at the time MSCAN did not support genomic data so that need to be stored into my separate database for analysis.

Data that comes from the group database is provided in CSV (comma separated value) format. Python includes a csv module that allows for fast script development for importing data, so a series of routines were developed for automating data import, including the use of python to access serialised versions of certain database tables from the groups database, which were provided as JSON objects.

Data which was supplied by different *Miscanthus* studies came most commonly as Excel spreadsheets, some of which needed to be modified by hand to format the data in way that could then be processed in python.

In order to automate some of the analysis of data the module rpy2 was utilised. rpy2 provides a Python interface to the statistical language R. This allowed direct use of some

of the functionality of R in Python, which was mainly utilised for the plotting of data such as phenotypic data versus genotypic data, allowing large numbers of variables to be automatically plotted.

Analyses were performed using several different computing platforms including a Debian based, 12 core, 192 Gb RAM server and IBERS HPC, a Sun Grid multi-node cluster consisting of 400+ cores of either Intel or AMD design for data management and handling. Data processing and analysis with a multicore data extraction method was developed for combining genomic and phenotype data using Python and SPRINT library in R package.

2.5.2 Genotype data

Genotype data generated from the GBS were presented as a complete set of markers with its associated genotype and allele. This leads to the creation of several tables to store the combination of allele calls and map information. A Python-based database was designed to handle both physical map and genetic map positions. As currently no physical map exists for *Miscanthus* instead this was used for storing the positions of *Miscanthus* markers on the *Sorghum* genome. All markers are stored in a table, with a second table storing the allele calls for each. Additional information such as SSR phylogeny classifications were also stored in the database, this data was retrieved from the group database.

2.5.3 Phenotype data

Several important phenotypes were recorded and analysed in this thesis. Each plant within the IBERS' Miscanthus germplasm collection has an accession which tells of its collection information and also a unique identifier which is linked in the group's

database so the history of each plant can be queried. Data was exported from the group database into the Python database, which was designed to facilitate faster data analysis by changing the data structures used in the group database. A common set of fields existed in all the phenotype data which consisted of the plants UID and location and plot where the observations were taken. If the trial consisted of several replicated genotypes (clone created by rhizome splitting), its replicate number, link to the genotype record and the date of the observations were also collected and stored in the database. Data collected were used to create data tables containing meteorological, genotypic and phenotypic data. For field data collection I developed an application designed to run on the Android platform, so that a small handheld device could be used for data collection, such as a mobile phone or tablet. I designed the app to store a field plan which could be downloaded from the group database, MSCAN. The application would then walk the user through the field requesting phenotypical data on each plant, which were initially hard coded into the application. A later version would download phenotypical measurement definitions from MSCAN directly. The application was coded to record certain predefined phenotypes, and made data collection in the field more user friendly that current methods being utilised by the breeding team.

2.5.4 Meteorological data

Meteorological data were collected from a variety of meteorological stations in and around the several *Miscanthus* trials. All meteorological stations were represented in the database, and these were then assigned to a trial, so requests for linked phenotype and environmental data could be processed automatically. Several attributes were measured on a daily basis and several additional measurements were estimated using meteorological functions described below. Given that most climate models predict water

shortages in the next few years, study on drought tolerance is becoming an important trait for many crop species. To quantify water stress, raw meteorological data was used to calculate soil moisture deficit using the standard FAO Penman-Monteith equation (Allen *et al*., 1998), The equation calculates potential water loss from evapotranspiration (**ET_o)** , as shown below:

$$ET_o = \frac{\Delta(R_n - G) + p_a c_p \left(\frac{e_s - e_a}{r_a}\right)}{\Delta + \gamma\left(1 + \frac{r_s}{r_a}\right)}$$

where

R_n is the net radiation,

G is the soil heat flux,

e_s is saturation vapour pressure (kPa)

e_a is actual vapour pressure (kPa)

p_a is the mean air density at constant pressure,

c_p is the specific heat of the air

Δ slope vapour pressure curve (kPa °C^{-1})

ɣ is the psychrometric constant (kPa °C^{-1})

r_s and r_a are the (bulk) surface and aerodynamic resistances.

r_a is calculated using the flowing formula

$$r_a = \frac{\ln\left(\frac{z_m - d}{z_{om}}\right)\ln\left(\frac{z_h - d}{z_{oh}}\right)}{k^2 u_z}$$

where

r_a aerodynamic resistance (s m^{-1})

z_m height of wind measurements (m),

z_h height of humidity measurements (m),

d zero plane displacement height (m),

z_{om} roughness length governing momentum transfer (m),

z_{oh} roughness length governing transfer of heat and vapour (m),

k von Karman's constant, 0.41,

u_z wind speed at height z(m s^{-1}).

r_s is then calculated as

$$r_s = \frac{r_i}{LAI_{active}}$$

where

r_l bulk stomatal resistance of the well-illuminated leaf (s m^{-1}),

LAI_{active} active (sunlit) leaf area index (m^2 (leaf area) m^{-2} (soil surface))

Note that a measure of leaf area index (LAI) is also required for Penman-Monteith equation for which the FAO (Food and Agriculture Organization of the United Nations) standard grass land numbers were used as an approximation. By combining the evapotranspiration loss with the rainfall and assuming the plot is a closed system which can only lose moisture via evapotranspiration, we can then work out the soil moisture deficit (SMD) for each plot. This is a simplified model but it can still work as an approximation of water available to each plant. SMD can then be combined with other

meteorological measurements to describe the environmental conditions in which plants were grown. Degree days were also calculated at several base temperatures (0, 6 10°C) using the McVicker's (McVicker, 1946) formula:

$$DD = \frac{(T_{min} + T_{max})}{2} - T_{base}$$

where

DD is degree days

T_{min} is the daily minimum temperature (°C)

T_{max} is the daily maximum temperature (°C)

T_{base} is the base temperature for the given crop.

A base temperature of 10°C is commonly used for *Miscanthus*.

PAR is also used in this models developed in this thesis. PAR is a measurement of the amount of radiation from the sun in which the wavelength lies within the range that can be utilised by photosynthetic organisms. This is often measured in μmol photons $m^{-2}s^{-1}$. The measurements from the PAR sensors have been converted into MJ per day using the assumed value of 1800μmol to 1J.

A parameterisation system was created to generate meteorological variables for the machine learning algorithms to detect the differences between years. To do this first a time period must be selected, in this thesis two have been used, 7 days and 5 days. Next the meteorological observations for use are selected. In this thesis different observations were used in various experiments, these included degree days, temperature (minimum and maximum daily), PAR (photosynthetically active radiation), rainfall and soil moisture deficit (SMD). Once the set of observations is selected the daily measurements for the year is

divided into subsets, each the length of the time period chosen. Depending on the time period selected the last subset may contain less observations.

Depending on the observation the daily values are then either summed or averaged in each subset. For PAR and rainfall summation was used, for all others an average was calculated. The sums and averages each become an attribute. So for example if a time period of 5 days is used there would be 73 (365 days per year over 5 days observation window) attributes for each observation. Therefore if minimum and maximum temperature and rainfall were include in the model using a 5 day observation period this would give a total of 219 (73 values * 3 observations) meteorological attributes into the model.

2.5.5 Data export and handling

To facilitate the machine learning and other data analysis, an export pipeline was created to link genotypic, environment and phenotypic data together. A Python based interface was created to allow for datasets to be defined and exported into CSV and ARFF (the relation file format used by WEKA) (Hall *et al*., 2009) formats. This involved the creation of a Python script in which the data set to be created is defined. A collection of Python objects were created which hold a set of parameters which defined what data is to be exported. These parameters let the user choose what data is exported, such as genetic, phenotypic or meteorological. It also included a filter system to allow selection of subsets of data, such as one particular field trial. Data were integrated from several projects and therefore required the standardisation of data before analysis.

As discussed earlier, the majority of genomic data used in this research was SNP marker data. A methodology has been developed to scan a large number of SNPs to find those in association with the trait being investigated. The machine learning approaches do

not require data to be normally distributed and can deal with noisy data sets like the data that come from phenotypic measurements where accuracy can be difficult to achieve. We used the R (R Core Team 2013) implementation of random forest (Liaw & Wiener 2002) and an R parallelisation package SPRINT (Hill *et al*., 2008) for data analysis.

Markers were recorded in the database as the observed allele from the GBS analysis. The values A and B were used to represent two possible homo-zygotes and the value H is used to represent heterozygotes. A fourth value '0' is introduced to code for missing allele values in some genotypes, which might be caused by either a missed read or the marker not being present in that genotype. All marker variables were assumed to have the same number of levels {A, B, H, 0}, irrespective of whether the allele was observed for that marker in the population.

A data matrix was created with each column being a marker and each row being a plant. If data were collected from multiple years or locations, or both, additional attributes were needed to allow the model to account for the cause of additional variance. If the plants were perennial the age of the plant was calculated from the year of planting, a plant being considered to have an age of 1 if the observation in question was taken in the year of planting. This is then added as another attribute. If data was taken from multiple locations and years, instead of just using a factor variable, as described above, meteorological variables can be used.

This Python export script was then run to create either CSV, ARFF of both types of data files that could then be used by WEKA or loaded into R.

2.6 Models Validation

One of the most important steps in model development is confirming that the model

created is accurate and is effective in describing the problem. This body of work applies machine learning to complex biological problems attempting to uncover relationships between genotype, phenotype and the environment in a species for which only limited genetic studies have been performed. This means that data for comparisons and validation can be difficult to obtain, therefore requiring model validation to come from several sources.

2.6.1 Validation methodologies

Machine learning has a wide array of algorithms and approaches that can be applied to the discovery of hypothesis and modelling of complex problems, although no one method is widely accepted, instead it is often a case of finding the best method for a given problem. Hence, with such a wide selection of available approaches, model validation must be used in order to discover the best algorithm for the current problem. Several different types of model validation can be used to achieve this.

Conceptual research deals with the task of verifying whether an analysis method creates a model accurately representing the problem in question. The validation of a conceptual research model requires knowledge of the problem or similar problems to be able to have data or hypotheses to confirm if the created model explains the nature of the true problem. In biological systems it is often the case that the exact nature of the problem is not known in a given organism, but it might be known in other species. Given the nature of evolution and relationship between species it is fair to assume that the behaviour in one species may be similar to others. For example studies have looked into the synteny between flowering time in Rice (*Oryza sativa*), Maize (*Zea mays* L.) and *Arabidopsis thaliana* (Cardon *et al*., 2004). Other studies have also shown that synteny exists between

various crop species (Armstead *et al*., 2004; Choi *et al*., 2004). These relationships between crop species mean that functions of some genes and genetic loci in one species can be potentially validated by utilizing studies carried out in a more extensively mapped species.

Working intensively on a single species, e.g. model plants such as *Arabidopsis* (Meinke *et al*., 1998), allows for a large amounts of discoveries to occur leading to a vast database of knowledge of genes, pathways, phenotypical responses, mutations etc. The idea is that this data can then be used to form hypotheses in other plants, based upon the observations in *Arabidopsis* or other more closely related model plant species. Comparison of SNP markers however is made more difficult by the fact that they are unlikely to have been identified using the same methodology; changes in enzymes used in digestion of DNA, differences in the bioinformatics pipeline and other changes can lead to the detection of slightly different markers. Therefore an exact marker match is unlikely unless the same datasets are used in both experiments and validation will usually come from markers clustering in similar genomic regions rather than by direct comparisons.

Although it is possible that many genes may be in linkage disequilibrium with a single SNP there are ways in which the related gene can be potentially identified. The use of RNA experiments, which measure gene expression in different tissues or growth stages, can be used to highlight potential genes that may be controlling the trait of interest.

Empirical research aims to develop knowledge by using direct or indirect observations of a problem. Often the researcher will have a hypothesis they wish to test. By utilising experimental design a researcher will create an experiment through which data can be collected to test their hypothesis. As with the conceptual studies, in crop science data sets can be found online from many crop species. They can be used to test modelling

approaches. Many SNP studies in more widely studied crop species are available. Rice is one example, with one study publicly releasing a 44,000 SNP dataset which was used for a GWAS analysis of many traits (Zhao *et al*., 2011). By the use of these data sets methods can be tested for accuracy and the ability to detect SNP markers related to these traits. This will allow for the modelling approach to be first validated in Rice before being applied to *Miscanthus* where validation of discoveries may be more difficult.

Eclectic research aims to collect a wide range of data on a problem and then use a more general approach to look for clues that will lead to the hypothesis discovery within this data and route research in the right direction. Described by Armstrong (Armstrong, 1974) as the shotgun approach to hunting, 'shooting' over a wider area could potentially reveal more information without necessarily leading to a direct 'kill', whereas intensive research is akin to a rifle, where a single shot either hits the target with a 'kill' or misses, revealing no more insight. To put it into a biological data context, firstly data may be collected on a large population with is fairly high level, such as recording the plants which look healthy in drought or hot conditions. Over time clusters of these reports can be created along with genomic data to look for differences that separate these groups, which can then be investigated more intensively, saving time and reducing the resources expended in discovery of the hypothesis.

The three research approaches discussed above have been used in scientific discovery, but each has its own advantages and disadvantages which must be accounted for when selecting a methodology.

Conceptual analysis allows for general models to be created by relying on relationships between species. Although even within plant species where there is a high level of synteny, for example Rice and *Arabidopsis*, there can be differences between the

control of various traits, such as long and short day length responses in flowering (Izawa, 2007). If a reference plant has evolved away from a model plant then a model developed for one might potentially not work for the other, or be less accurate. Hence many different model plants are being developed. There are places where conceptual models do have value though - the genome. The long chain of ATGC which define the blueprint of life is common to all living organisms, and so development of a conceptual model at that level would inform all research, by understanding what controls gene expression and other similar phenomena. One advantage of concept based models is that they can be developed for problems which might not actually be able to be tested empirically and then tested against known examples in order to see if the logic fits. Although this does not validate a theory it may show it as a best fit given current knowledge. Conceptual models are useful in science but their scope may be limited depending upon the underlying assumptions. Models that are too specific might not be able to be applied to other problems.

Empirical research focuses on developing models for a predefined hypothesis that is able to be replicated and through its design can be considered to be valid. Strict experimental design is often applied in this context, by fixing variables and altering only those of interest in an attempt to eliminate all irrelevant variance. This is done to allow the researcher to validate the hypothesis they wish to test. This type of experimentation aims to reject or support a hypothesis. However it can not provide proof of a hypothesis. If the experiment does not lead to a rejection of the hypothesis, it can be considered valid under the conditions it was tested. Empirical research does not perform hypothesis discovery directly so is limited in its scope. However, its support or rejection of a working hypothesis can lead to new ones being formed. Unlike conceptual analysis empirical research aims to

answer a single question, not develop a concept that can then be applied to other areas of research.

Last is the eclectic research approach, the idea of using generic exploration data to discover a hypothesis or clues that would lead to the development of more specific hypotheses. This approach comes with risks. When analysing the data the 'general' hypotheses could be missed or the theories developed could lead the analysis in an incorrect direction of false hypotheses or in which no hypotheses exist. Although if the collection of general data can be performed cheaply, then it may lead to correct idea reducing costs compared to that of the other methods. In essence this is what is done in data driven science - large volumes of data on the general concept is analysed to look for patterns that can lead to the development of hypotheses. This type of data is often cheap to acquire. Within the business sector this data maybe in the form of web traffic, server logs, and transaction records. All these data are collected as part of the normal behaviour of a website so come at no extra cost. Businesses wish to then utilise this data to develop models of customer behaviour.

All three approaches have their place in scientific discovery - learning concepts allows for understanding of the nature of common elements, empirical tests can validate particular hypotheses, and eclectic research can be used to discover concepts and hypotheses from many data sources which can then be taken onto further testing or for use in other studies.

All three of these methods can be used in conjunction, which presents a powerful methodology for data discovery, if we take the example of a plant breeding program. A breeding program may collect large volumes of observations on plants, these potentially maybe fuzzy labels, such as drought tolerance, good performance in high temperatures.

Using eclectic methodologies this data could be processed to look for potential sub populations with similar behaviours. This information could then be used to create a concept of say drought tolerance, possibly using observed phenotypes e.g. plants with thick leaves are drought tolerant. This could then be used to select plants for drought tolerance when selecting parents, meanwhile a subset of these plants could then be empirically tested in order to discover if they really are drought tolerant.

This research will use the eclectic theory, by using data available on many progeny from historical observations and meta data such as planting times, taken from the group database (MSCAN). The aim is to develop models that can be used to aid breeding by the fast and simple detection of markers that relate to phenotypes of interest. Other studies will look at understanding phenotypic traits by using large volumes of data from various meteorological stations and other sources.

The above deals with how research in general, and research within the thesis in particular, is performed and evaluated. The following will look at statistical methods for evaluating the 'goodness' of a created model.

One of the most classic ways to test the ability of a model is to measure the mean squared error, the sum of the squared differences between predicted and observed values. In order to test the model’s ability to generalise on the problem one can calculate mean squared error using previously unseen data, which had not been used in fitting the model, often referred to as a validation data set to work on previously unseen data. Therefore testing its ability to generalise on the problem. Mean squared error (MSE) is often used to measure a model’s performance, and is calculated using

$$MSE=\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i-y_i)^2$$

which is the average of the sum of the residuals from the model. Alternative measures of error include root mean squared error (RMSE) which is the root of the MSE.

$$RMSE = \sqrt{\frac{\sum_{i-1}^{n} (\hat{y}_i - y_i)}{n}}$$

Many more approaches exist to measure error including mean absolute error. In essence all of these methods simply take the sum of the difference between predicted and actual values and use some transformation to turn all results positive. Both of these methods can be used to test the accuracy of a fit between two different models.

R^2 provides a measure of how well a model performs against a model that uses an observed mean to predict all values. R^2 represents the 'proportion of variance' explained by a model. If R^2 is close to 0 then the model can be considered to be only performing slightly better than just selecting the mean value. However the R^2 as presented in linear regression is the square of the correlation between predicted and observed data points. As with MSE or RMSE this can be performed on both validation or test sets, and R^2 is almost like a MSE scaled by data variance. R^2 is calculated like so:

$$R^2 \equiv 1 - \frac{SS_{RES}}{SS_{TOT}}$$
$$where:$$
$$SS_{TOT} = \sum_i (y_i - \bar{y})^2$$
$$SS_{RES} = \sum_i (y_i - \hat{y}_i)^2$$

The above are all simple methods for testing model error. Of course there are more complex methods that can be used to test a model's ability, which provide better testing for generalisation.

Cross validation is one method to test an algorithm's ability to develop a model that is

capable of generalising its understanding of a problem. *K*-fold validation is one classic example of cross validation. This approach first splits data into *k* number of subsets, then one approach is to remove a single subset from the total data to be used as a test set. The model is then created on *k* -1 subsets, and tested against the missing subset. This is repeated *k* times, removing a different data set each time A special form of cross validation is known as leave-one-out. In this model k = n, where n is the number of observations. The model is fit on n-1 observations and then tested against the single removed observation. Once the cross-validations predictions are performed they are then tested against the observed data.

There are other considerations when creating a model, one of the most important is interpretability. Being able to understand what the model is doing/has done/does is especially important in experiments where modelling is being used to perform knowledge discovery as is often the case when applying machine learning to high dimensional data sets. The goal is to find the influential factors and learn how they interact to create an understanding of the problem so that a hypothesis can be formed.

Not all machine learning or statistical methods are equal when it comes to interpretability. Take artificial neural networks (ANN) and decision tree learning for example. Details of implementations of these two models will not be considered - only the resulting models and the interpretation of each. As discussed earlier, ANN’s are created out of a highly interconnected network of model neurons which consist of multiple layers and weights which control the signals between nodes. ANN's could be used as a classic example of the black box model in which data goes in and results come out but what happens in the network is far from clear. Of course one could follow a signal propagation from data through the network and see how the learned weights cause various areas to

'trigger', but this does not really explain the decisions taken by the learner.

Decision trees on the other hand results in a collection of binary split nodes in which each 'split' in a parameter can be seen and it is very easy to follow the tree from root to leaf node and understand the decisions the learner has made. Even more complicated versions of decision trees, such as random forest which creates many trees, can be interpreted via the importance scores given to each parameter.

Although this example looks at two extremes of model interpretability in machine learning, it does highlight another point about method selection and how that can influence the way models can be evaluated. Attempting to conceptually evaluate a model without being able to understand how it works is much more difficult.

In biological studies where one is not so much concerned with a prediction but more with understanding of how, for example, a plant's genotype is related to its flowering time, a model that is very 'human-readable' is of critical importance. The trade off between readability and accuracy depends on how the model is to be used. For scientific discovery, the understanding of what the model means maybe more important than accuracy. However in breeding where the goal is to get the best progeny accuracy maybe more important that understanding. So model validation can be different depending on the problem, but if a model could be found that provides both good understanding of the problem and accurate predictions then both disciplines can gain benefit from the same experiments.

The validation of a model is clearly a complex question which depends on many factors. Some of these factors can be simply expressed in metrics, such as MSE or RMSE, which clearly define which model is better, or R^2 which gives an idea of how well a model is understanding the trend in the data. Approaches such as cross validation allow

for better testing of models and confirmation of generalisation by utilising the data many times by adjusting the instances seen in the data set. The metrics previously mentioned, such as MSE or R^2 are also often used to evaluate the results of a cross validation.

Validation of a model is the key stage in any modelling experiment. Selection and testing are important but if the model cannot be validated it cannot be considered to be useful. How a model will be validated is a consideration that must be first thought of before analysis can begin. Bad model selection at the early stages can potentially lead to a model which cannot be validated conceptually or empirically, rendering the model useless.

In this research the dataset is randomly split into two data sets, one for training and the other as validation data. The split should favour the training set, as more data for the algorithm to formulate the rule will result in a better model. Most commonly used is 85% for training and 15% for validation, but other splits are possible. The model will only use the training data set and the validation set should be set aside for later use. Random forest function has been applied to the training dataset and the trait of interest is used as the response variable in supervised mode. Depending on the response type the random forest function will automatically select regression or classification. Random forest's importance scores will give a higher score to classes which have more levels, this is not a problem in the marker data sets as all markers will only have the four levels A, B, H and 0 as outlined earlier, so corrections on importance scores will not need to be performed. The SPRINT package allows for parallel implementation of the random forest function which helps to reduce computation time of the random forest tree construction.

All the traits we modelled using random forest were treated as continuous variables. The number of trees to be grown in the forest was 500, this allowed for the error rate to plateau while still allowing for a reasonable computing time. Across multiple different data

sets 500 trees was found to be more than enough for error rate to minimize. Prediction accuracy of the results can be calculated by comparing predicted values to the actual observations by calculating the correlation between the two vectors. Depending on the prediction accuracy achieved, it may be necessary to tweak some tuning parameters, like the number of trees or the number of variables randomly selected to build each tree, and repeat the whole process. Once satisfactory prediction accuracy was reached, the validation process was started. The prediction results were compared against the actual observations in the validation set.

Once the final model was chosen and assessed, a process of examining the markers selected by the model can proceed. We do this by studying markers' importance scores which are used to rank parameters in order of their influence on response prediction. The importance score for each parameter can be extracted using a python script which uses attributes from the export script with a collection of graphical generation functions that show importance score plotted against any available genetic or physical map. Once important markers have been identified we can then look at were they are located in the genome. This will allow for investigation of potential genes or QTL that occupy the same regions. If an organism has the whole genome sequenced and annotated, this annotation can be used to search for potential genes that control particular trait. Otherwise a reference genome can be used.

Markers with the highest importance score were used to locate potential genes or QTLs that lie within a given region. All genes within 100kb of either side of the markers were investigated. If additional data such as expression data or gene ontologies were available these were used to look for potential genes that could be related to the trait in question. Where proteins for related genes were known PFAM (Punta et al. 2012) and

Interpro (Hunter et al. 2012) were also queried to look for any potential link.

2.6.2 Methodology Testing

In order to test the ability of random forest to detect marker-trait associations it was decided to use a publicly available data set of a more widely studied crop than *Miscanthus*. It was also advantageous that this dataset had undergone some quantitative genetic analysis so that comparisons between methods could be drawn. To do this a genome-wide association study in Rice was selected (Zhao *et al.,* 2011). Zhao's study used a 44k SNP chip to generate marker data for 400 genotypes of rice which also had phenotype data for several traits. It was decided to randomly select a third of the SNP's from this dataset in order to reduce the computation time. This left 12302 markers for the random forest model to analyse.

In Zhao’s experiment the flowering time observations were taken in Aberdeen in the form of the day of year (DOY). The rice annotated genome (RAP-DB) was used to search for the genes within the 100kb range of some of the significant SNPs (Sakai *et al.,* 2013). RNA data from 7 tissues, PABF, panicle before flowering; PAAF (panicle after flowering); RO (root); SE (seed); SH (shoot); CA (callus); LE (leaf) were captured using RNA-Seq (Sakai *et al.*, 2011). Reads from the sequencer are given as FASTQ files, these contain the RNA-Seq data. The reads were then aligned to the reference genome from RAP-DB using the software tool TopHat (Kim *et al.*, 2013). Then the software cufflinks (Trapnell *et al.*, 2010) was used calculate a FPKM (Fragments Per Kilobase of transcript per Million mapped reads) value, a measure of relative transcription, for each gene in the annotated genome. An additional dataset was also acquired that contained the known QTL for flowering time in rice (Yonemaru *et al.*, 2010) (Table 2.5). Interpro was used to check for the homologs of the proteins coded for by genes where the function was unclear and was

to be checked against other species (Hunter *et al.*, 2012).

The results using the random forest methodology are presented as a graph of the actual observations versus the predictions created by the model which is based on a subset of the markers (Figure 2.5). An R-squared of 0.32 is seen in this dataset, but the variance explained by the model was 31.86%. This is less than found by the GWAS model, but recall that only a third of the markers were used to reduce computational time.

The importance scores from the random forest model for Aberdeen flowering time were plotted against the physical location on the genome (Figure 2.6). The results indicated that there is a region of chromosome 6 that is highly important in flowering time in rice. The top 50 significant markers with their locations and importance score can be seen in Table 2.6. The top 6 markers all lie on chromosome 6 and reveal four significant peaks within a similar region. The most important one appears at position 8,183,433bp. Phenotypes for the alleles of this marker appear to show that the B allele seem to lead to a later flowering time (Figure 2.7). The fourth most important marker appears approximately 3000 bp away from the top most important markers. The second most important appears at position 7,923,685bp. The third most important appears at position 9,362,972bp while the fifth and sixth appear at positions 9,834,592bp and 10,000,715bp, respectively. Thus, the top 6 markers appear in a range of approximately 2.1Mbp within each other and since the length of chromosome 6 is 31Mbp this means they lie within 6.7% of the length of chromosome 6.

The 7th most important marker appears on chromosome 4. The phenotype plotted against the alleles of this markers shows a slight association of the A allele to later flowering phenotype (Figure 2.8). On this chromosome there is only one other marker showing importance, the 12th marker which lies only 5kb away. The 9th marker appears

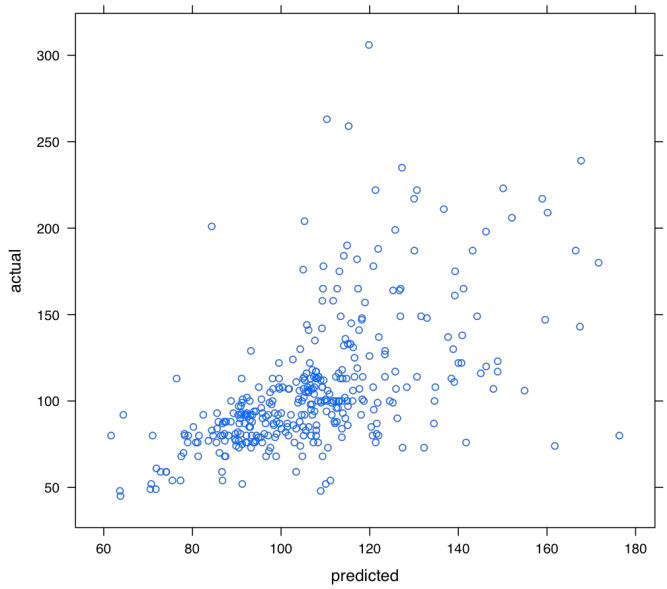

Figure 2.5: The predicted values for flowering DOY for rice at Aberdeen from the random forest model versus he observations taken in Zhao's study. We see that for the early flowering plants the random forest predicted well.

on chromosome 1 but it is the only marker in that locus which shows any importance for flowering. When the phenotype for the alleles of this marker were examined it showed that lack of this marker or the presence of B allele seem to associate with a later flowering (Figure 2.9).

The rice GWAS study made use of all 44k markers and, as mentioned before only a subset was used in our random forest analysis. Hence we only used the top ranked markers from the rice GWAS study that appear in this subset.

A list of significant SNPs from the GWAS study was compiled. The total documented number of significant SNPs was 97, among them 37 appeared within one third of those selected for analysis using random forest. Significant SNPs from the GWAS study were compared to the top 50 markers selected by the random forest analysis, with the penultimate column of Table 2.6 indicating whether each marker is significant in the GWAS study. 51% of SNPs (19 out of 37) in the 1/3 subset used for our analysis that were

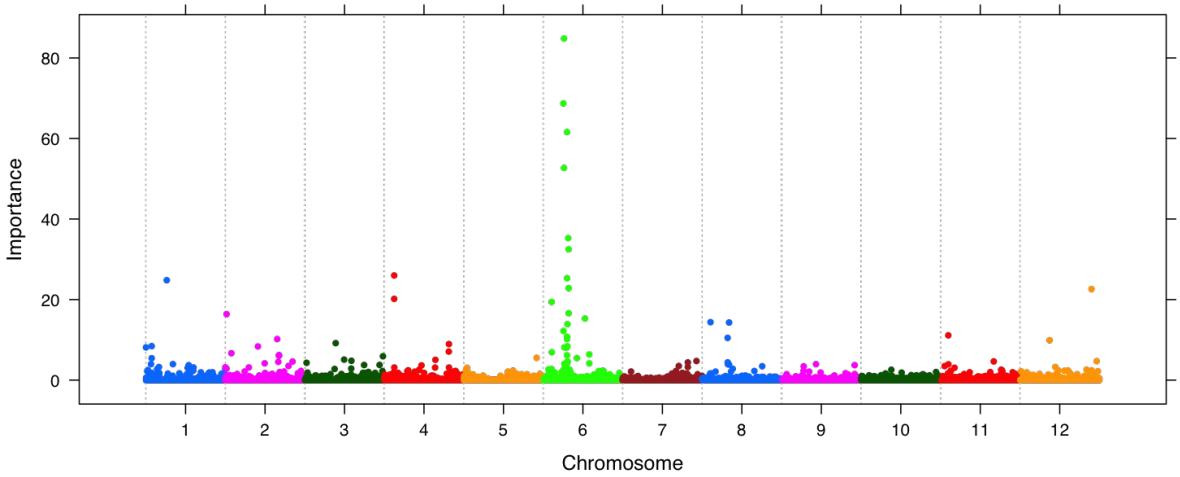

Figure 2.6: The importance scores from the random forest analysis of flowering time in rice from the Aberdeen data. We see a peak of significance on chromosome 6, and two smaller peaks appear either side of this. Also a peak exists on 8th, 4th, and 12th chromosome although these are not as high in importance and appear to be only a small number of markers, whereas the 6th chromosome group consists of many markers.

highlighted as significant by GWAS also appear in the top 50 of the marker pool chosen by the random forest model. These include the top 6 markers on chromosome 6 that were discussed earlier. Out of the top ten random forest selected markers those on the 4th and the 1st chromosomes do not appear significant in the GWAS study. As mentioned earlier the marker on chromosome 1 has no other important markers nearby which could imply this is a false positive, and maybe the SNP has no effect on the trait, but could be in linkage with another loci that has significant effect. We also compared the top 50 random forest SNPs against 49 known flowering QTLs from the Q-TARO database (Yonemaru, 2010) (shown in Table 2.5). The last column of Table 2.6 lists any flowering QTLs that lie within 100kb either side of each random forest marker. In total 27 out of the top 50 SNPs have at least one flowering QTL within their 200kb catchment area. The top 6 SNPs, in particular, all lie within the 100kb range of qHD-6-1 and/or Unamed 2, both of which are QTLs linked to flowering time.

Figure 2.10 shows importance scores of all the markers used in random forest analysis plotted against their chromosome position with QTL positions superimposed.

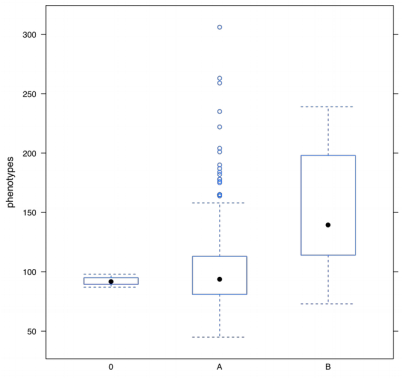

Figure 2.7: Flowering DOY distribution show for the top marker selected by random forest, which appears on chromosome 6.

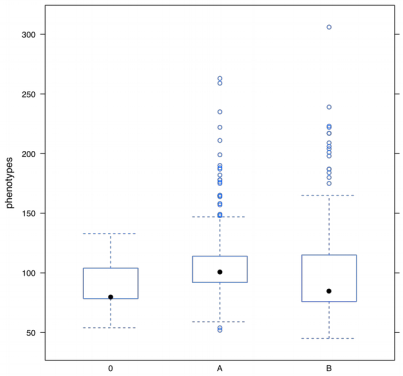

Figure 2.8: Flowering DOY versus allele for the marker which is 7th most important as selected by random forest, this appears on chromosome 4.

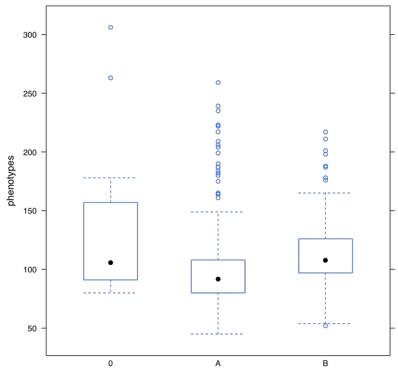

Figure 2.9: Flowering DOY versus the alleles for the marker which is 9th most important as selected by random forest, this appears on Chromosome 1.

Importance scores are shown as vertical lines so intersection with QTLs can be seen. Analysis has shown many areas of high importance are within or close to known flowering QTL. However, there are several high importance peaks that are not near any known flowering QTLs, most notably on Chromosomes 1, 4 and 12. They are related to SNPs which are ranked 9th,7th and 11th by the random forest, respectively. Instead, in these cases RNA expression data were used to search for possible candidate genes that relate to flowering.

The 9th SNP (id1008137) is located at 11376832 on chromosome 1. Using the FPKM values from the cufflinks analysis mentioned earlier, scans were made within 100kb of either side of this SNP and genes which have highest expression in the panicle, before and after flowering were selected. The expression values for these genes can be found in Figure 2.11.

There are five genes that lie within 100kb of the SNPs that had higher expression in the panicle than in the other tissues. One gene, Os01g0306200, shows high levels of expression in the panicle when compared with other tissues measured. This gene is recorded as producing a protein of unknown function DUF3511 domain containing protein.

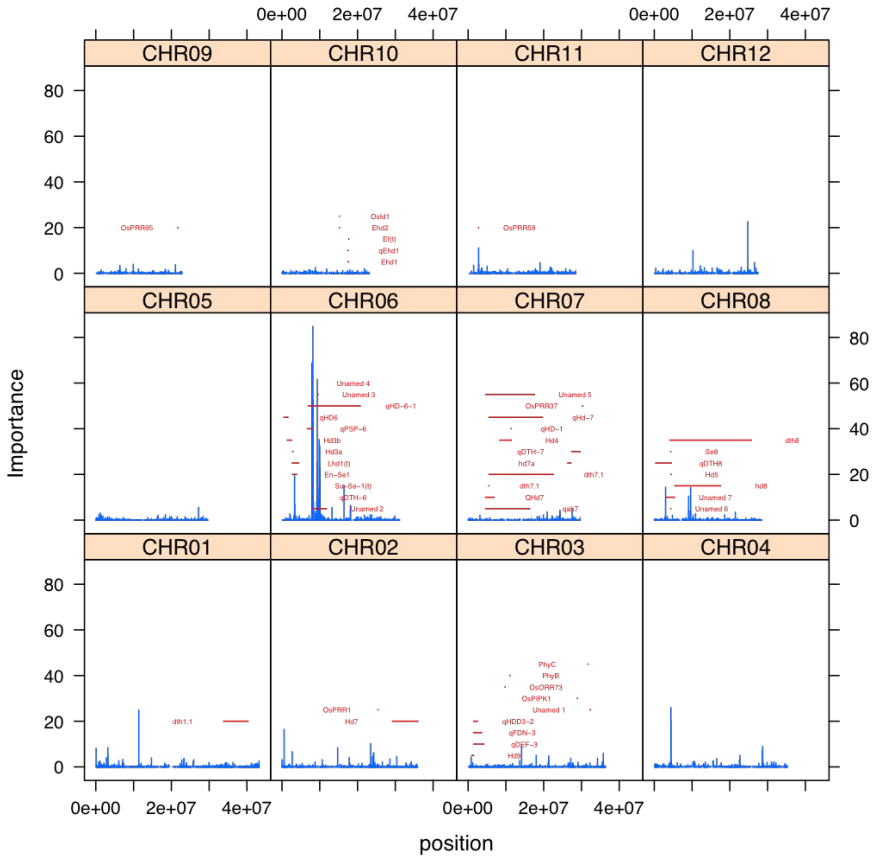

Figure 2.10: Importance score from the random forest analysis are replotted as vertical lines, and the QTLs from the Q-TARO database are plotted to highlight where a region of importance match to known QTLs for flowering in rice.

Interpro results for DUF3511 revealed 274 matched proteins and matches were found in a range of flowering plants such as *Pinus radiata* (Monterey pine), *Brachypodium distachyon*, *Lotus japonicus* and many other plant species. This protein has only been found in flowering plants which could indicate its use in plant only functions. As it has been shown to be near a SNP that has been marked as important for flowering it could be that this gene is in fact involved in flowering initiation possibly explaining why it is only present in flowering plant species.

The 7th SNP (id4001850) is located at 4409758 on chromosome 4. The FPKM values using cufflinks are illustrated in Figure 2.12. Seven genes appear within the 200kb window

QTL/Gene	Chr	Genome start	Genome end	LOD	Character	Explained var.	Additive effect	Year
dth1.1	1	33817920	40288357	3.42	days to heading	7.22	-	2008
Hd7	2	29242392	35959447	3.2	heading date	13.8	2.4	2000
OsPRR1	2	25427321	25430309	-	relative amount of mRNA	-	-	2003
Hd9	3	975995	1427051	-	days to heading	-	-	2002
qDEF-3	3	1423343	4098191	4.7	days to emergence of flag-leaf	10.9	-3.17	2004
qFDN-3	3	1429107	3509693	4.19	Flowering duration	18.3	3.44	2002
qHDD3-2	3	1429107	2432615	8.32	Heading date	32.1	4.35	2003
Unamed 1	3	32239082	32365157	21.3	Days-to-heading	33.3	6.4	2007
OsPIPK1	3	28940657	28946571	-	heading date	-	-	2004
OsORR73	3	9810142	9819353	-	relative amount of mRNA	-	-	2003
PhyB	3	11070754	11078864	-	-	-	-	2005
PhyC	3	31767880	31772937	-	-	-	-	2005
Unamed 2	6	8054255	11750090	16.25	heading date	0.42	9.45	2005
qDTH-6	6	8054255	8066362	-	days-to-heading	34.6	-7.9	2001
Su-Se-1(t)	6	8054255	8066362	-	dominant photoperiod-sensitive suppression gene	-	-	2005
En-Se1	6	2770072	3826329	-	days to heading	-	-	2000
Lhd1(t)	6	2684129	4352379	-	-	-	-	2000
Hd3a	6	2839864	2912060	-	days to heading	-	-	2002
Hd3b	6	1370829	2501902	-	days to heading	-	-	2002
qPSP-6	6	6720901	8066362	-	photoperiod sensitive phase	8.4	-	1998
qHD6	6	483009	1562787	4.26	Heading date	13.9	2.71	2006
qHD-6-1	6	6927624	20691040	3	heading date	7.3	-	2004
Unamed 3	6	9536259	9537572	-	-	-	-	1990
Unamed 4	6	8054255	8066362	5.1	Heading date	27.6	-	2001
qah7	7	4606397	16264722	3.11	length of the heading period	14	-	2004
QHd7	7	4606397	6812968	12.2	heading date	-	4.1	2003
dth7.1	7	5512628	5512754	4.95	Days to heading	12.2	2.54	2003
dth7.1	7	5512628	22532504	16.72	Days to heading	25.8	-	2003
hd7a	7	26313662	27191049	5.8	Heading date	11.1	3.77	2002
qDTH-7	7	27391198	29608218	28	day to heading	64.5	7.1	2005
Hd4	7	8358800	11394315	-	days to heading	-	-	2003
qHD-1	7	11391449	11394315	11.2	days to heading	38	-	2007
qHd-7	7	5512628	19619933	8.4	heading date	15.14	5.2	2007
OsPRR37	7	30276864	30289374	-	relative amount of mRNA	-	-	2003
Unamed 5	7	4606397	17535483	52.3	heading date	69.3	-13.1	2008
Unamed 6	8	4377457	4377597	-	days to heading	13	4.6	2004
Unamed 7	8	3170545	5333855	4.43	heading date	22.9	-	2002
hd8	8	5421262	17528755	7.2	heading date	23.1	6.25	1995
Hd5	8	4444681	4446617	-	days to heading	-	-	2003
qDTH8	8	360155	4446617	22.65	Days-to-heading	22.92	9.25	2007
Se8	8	4377457	4377597	-	Days-to-heading	29	54.2	2007
dth8	8	4105519	25684949	27.66	days to heading	51.1	9.31	1996
OsPRR95	9	21716815	21721485	-	relative amount of mRNA	-	-	2003
Ehd1	10	17481862	17627660	-	Early heading date	-	-	2004
qEhd1	10	17505263	17510657	10.3	-	50	2.1	2001
Ef(t)	10	17684573	17686581	-	heading date	-	-	1998
Ehd2	10	15197103	15199951	-	heading date	-	-	2008
Osld1	10	15197103	15199951	-	-	-	-	2008
OsPRR59	11	2772223	2776940	-	relative amount of mRNA	-	-	2003

Table 2.5: QTL's for flowering in Rice from the Q-TARO database

Rank	Marker	Importance	Chromosome	Position	Significant in GWAS	QTARO QTL Matches
1	id6005318	84.831300576	CHR06	8183433	1	in Unamed 2
2	id6004987	68.7199567176	CHR06	7923685	1	in qHD-6-1
3	id6005996	61.5919098997	CHR06	9362972	1	in Unamed 2
4	id6005309	52.6897068015	CHR06	8180085	1	in Unamed 2
5	id6006212	35.255257509	CHR06	9834592	1	in Unamed 2
6	ud6000461	32.4947089885	CHR06	10000715	1	in Unamed 2
7	id4001850	25.9901991275	CHR04	4409758	0	No Match
8	id6006031	25.3115115736	CHR06	9370687	1	in Unamed 2
9	id1008137	24.8101503645	CHR01	11376832	0	No Match
10	id6006256	22.8437593096	CHR06	10014027	1	in Unamed 2
11	id12008904	22.6107755158	CHR12	24744118	0	No Match
12	id4001817	20.1806928459	CHR04	4404736	0	No Match
13	id6002750	19.4147038812	CHR06	3330720	0	in En-Se1
14	id6006268	16.612001324	CHR06	10016072	1	in Unamed 2
15	id2000397	16.3806108296	CHR02	564804	0	No Match
16	id6009335	15.296156829	CHR06	16416489	1	in qHD-6-1
17	id8000908	14.3930907515	CHR08	3001212	0	in qDTH8
18	ud8000529	14.2990571486	CHR08	9629652	0	in dth8
19	id6006089	13.8869939607	CHR06	9530655	1	in Unamed 2
20	id6004993	12.20798038	CHR06	7955480	1	before Su-Se-1(t)
21	id11000873	11.1154163137	CHR11	2787119	0	after OsPRR59
22	id6005941	10.7860451186	CHR06	9339901	1	in Unamed 2
23	id8002877	10.4875965315	CHR08	9106121	0	in dth8
24	id2009403	10.1969476886	CHR02	23460330	0	No Match
25	id6006005	10.1837147205	CHR06	9366336	1	in Unamed 2
26	ud12000654	9.9043110136	CHR12	10245978	0	No Match
27	id3007092	9.1860617032	CHR03	14155877	0	No Match
28	id4009552	8.9556425413	CHR04	28696927	1	No Match
29	id1002518	8.4419900276	CHR01	3180818	0	No Match
30	id6006083	8.4082808897	CHR06	9529974	1	in Unamed 2
31	id2006052	8.365229346	CHR02	14740939	0	No Match
32	id1000003	8.1340488453	CHR01	73192	0	No Match
33	id6005402	8.1271130278	CHR06	8317786	1	in Unamed 2
34	id6005814	8.0813028187	CHR06	9203899	1	in Unamed 2
35	id4009507	7.0917161867	CHR04	28638322	0	No Match
36	id6002786	6.9110034896	CHR06	3381621	0	in En-Se1
37	id2001530	6.6823609496	CHR02	2734528	0	No Match
38	id6010122	6.3708043477	CHR06	18119356	0	in qHD-6-1
39	id2010176	6.2231342627	CHR02	24257914	0	No Match
40	ud6000446	6.1819141813	CHR06	9374625	1	in Unamed 2
41	id2010191	6.0411418612	CHR02	24262681	0	No Match
42	id3017884	5.9243277494	CHR03	35784000	0	No Match
43	id5013215	5.5325597874	CHR05	27199778	0	No Match
44	id6007954	5.473366826	CHR06	13248709	0	in qHD-6-1
45	id1002529	5.4387699265	CHR01	3183395	0	No Match
46	id3008808	5.0924226695	CHR03	18014678	0	No Match
47	id4007591	5.0362726723	CHR04	22683989	0	No Match
48	id3009980	4.7984129104	CHR03	21341625	0	No Match
49	id7005417	4.7463350878	CHR07	27547556	0	in qDTH-7
50	id12009816	4.7102560143	CHR12	26525678	0	No Match

Table 2.6: Top 50 markers for the random forest analysis of Aberdeen flowering

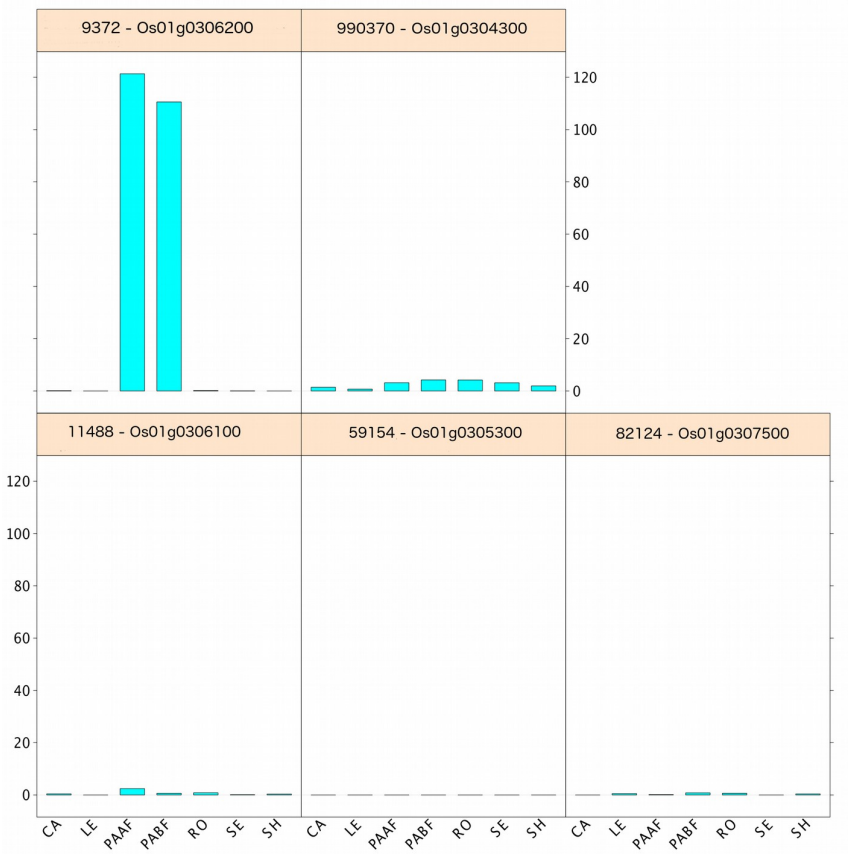

Figure 2.11: Expression values for genes within 100kb of the 9th most important SNP from the random forest model for flowering time in Aberdeen. Genes are only shown if their expression is higher in the two panicle tissue when compared with the other tissues observed, We see one gene has very high expression in the panicle both before and after flowering.

around id4001850. Generally the difference between expression in different tissues is less obvious than seen in Figure 2.11. However, we do see one gene where the PABF (panicle before flowering) expression is greater than any of the other tissues, Os04g0162600. Again, this gene produces a protein of unknown function, DUF295. When looking at the rice genome browser (Kawahara et al. 2013) the predicted gene Os04g0162600 appears to be overlapping LOC_Os04g08070. Orthologs of this gene exist in *Sorghum* and *Brachypodium* which both have been shown to create the protein DUF295. DUF295 has the Interpro reference IPR005174 and seems to match an F-Box containing protein seen in *Arabidopsis*. F-Box has been shown to be involved in gene regulation by providing

Figure 2.12: RNA expression values for all the genes within 100Kb of the 7th most important SNP from the random forest model. Its appears on chromosome 4. Although the differential expression for flowering tissues is not as significant as the SNP on chromosome 1 we still see a couple of genes that have higher expression in flowering tissues.

feedback loops (Shahri & Tahir, 2014). Random forest has the potential due to its data representation to detect associations that cannot be seen in conventional GWAS analysis. Given that this SNP has not been detected in other studies but was by the random forest, and given the potential role of F-Box proteins in gene regulation this gene could potentially have an epistatic relationship to the trait. Further study would be needed to confirm this.

The 11th SNP (id12008904) is located at 24744118. The RNA expression for genes within 100KBase of either side of this SNP can be seen in Figure 2.13. Five genes in that window had higher expression in the flowering tissues.

The use of random forest to predict flowering time in Rice did reveal similar results to that seen in the GWAS study, 19 out of a possible 37 SNPs selected by GWAS appeared in the top 50 hits for random forest. These top 50 also contain results that had not been

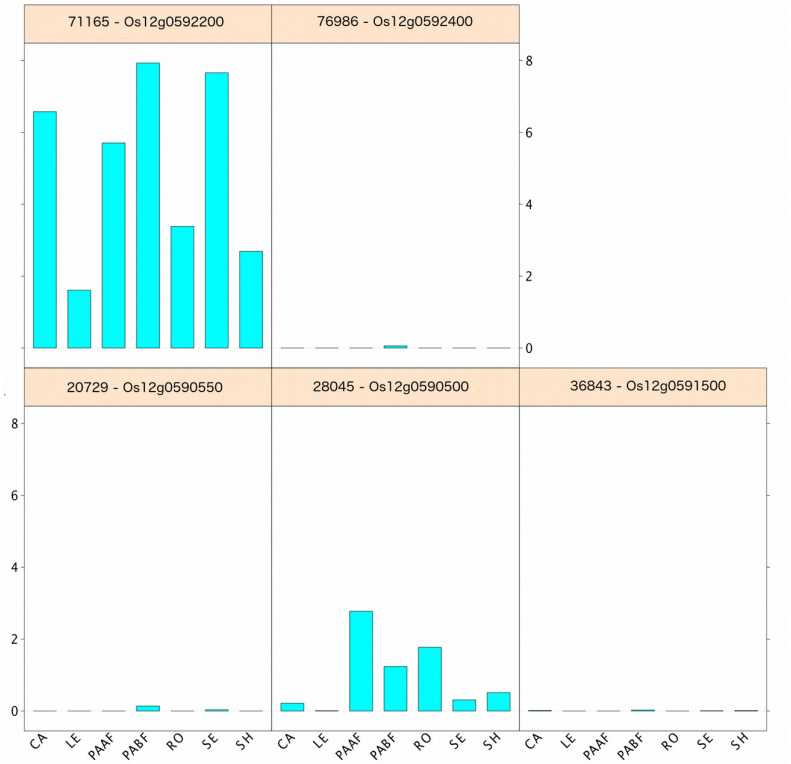

Figure 2.13: The RNA expression for genes that lie within 100Kb of the 11th most important SNP which is found on chromosome 12. Only 5 genes that have a higher expression in flowering related tissues exist in this region.

detected in the rice GWAS study.

From this exercise we can conclude that random forest is capable of not only detecting some of the SNPs highlighted by the rice GWAS study results, but also detect more QTLs involved in rice flowering time. Moreover, some of these new detected QTLs are located near genes that could potentially be involved in regulating flowering. Additionally random forest does not need kinship data, unlike GWAS studies, and needs only genetic marker calls.

3 Quantitative Trait Loci (QTL) Analysis using Machine Learning Approach

3.1 Introduction

QTL mapping is a widely used method for associating phenotypic variances with genetic markers and is used in breeding programmes to perform marker assisted selection. It demands a lot of effort in both phenotyping, generating and processing of genetic data. With ground breaking high-throughput technologies in genomics generating massive numbers of genetic markers, the conventional QTL analysis software such as MapQTL (Van Ooijen, 2004) needs a long computation time to handle such huge quantity of markers. MapQTL, for example, was unable to process the whole genomic dataset used in this chapter, methods had to be developed to reduce the number of SNP's used for each mapping. Recent genomic studies usually involve tens of thousands of markers (Tian *et al.*, 2011; Zhao *et al.*, 2011), therefore more computationally effective QTL analysis methods are needed to address those issues faced in the era of big-data biology.

Machine learning is a subfield of artificial intelligence. It uses an algorithmic model to detect patterns in complex data sets to provide predictions and selection of attributes that influence the result. Through building a model using supervised learning, the machine learning approach can teach an algorithm to find the relationship between the genotypic and phenotypic variances within a large problem space.

Miscanthus, an important energy crop, has been studied for phenotypic traits which effect yield (Robson *et al*. 2013; Jensen *et al*. 2013), and consequently optimising yields is the major goal in the breeding of *Miscanthus* as an energy crop. Studies have shown that

flowering has an extensive effect on yield (Gonza *et al*., 2001; Jensen *et al.,* 2013); and it has been shown to be a highly heritable trait (Slavov *et al*., 2014). In order to understand the genetic control of flowering time a mapping family, Mx2, was created at IBERS. Mx2 consists of 236 progeny from a cross of a late flowering *M. sinensis* and an early flowering *M. sinensis,* of which 185 were genotyped.

The primary objective of this chapter is to present the development of a generic and effective machine learning based QTL analysis tool to improve the QTL analysis and address the weakness of existing QTL analysis methods. Comparisons between this newly developed tool and existing QTL analysis software, including MapQTL (Ooijen, 2004) and SNP & Variation Suite v7 (Golden Helix, Inc., Bozeman, MT, www.goldenhelix.com), have also been performed for validation. With this section as introduction, the chapter contains five sections. Section 3.2 gives an overview of the Quantitative Traits Loci (QTL) and the technique used to detect the QTLs. Its role in molecular breeding and the weakness of current methods are also discussed in detail. Section 3.3 is devoted to presenting the methodology behind the newly developed machine learning-based QTL analysis tool. Section 3.4 presents the results produced from the application of this generic QTL analytic tool on energy crop *Miscanthus*. The flowering time dataset generated from the Mx2 mapping population is used to demonstrate the capability and strength of this tool. The results are extensively discussed and compared with results generated from existing QTL analysis techniques for validation. The chapter is subsequently concluded in Section 3.5 with a brief discussion.

3.2 Quantitative Trait Loci (QTLs)

Quantitative trait loci are the regions of a genome known to have an effect on a given

phenotypic trait (Kearsey, 1998). These regions often contain a gene which have an effect on the trait being modelled and the goal is to find and quantify its additive effect, a gain or loss to the phenotype being modelled. QTL have long been studied as a potential way to link genetic information from markers to phenotypic variances and have been widely used in the plant sciences with one study suggesting > 10,000 markers trait associations in various plant species have been reported (Bernardo, 2008). QTL mapping requires the creation of a mapping population with large scale of phenotyping, genotyping and the creation of a genetic map. It has been widely used in plant breeding to improve the effectiveness of selection by adding genetic information into standard breeding selection formula to utilise the genetic information to assist in breeding.

3.2.1 QTL and its role in molecular breeding

Traditionally, breeders use phenotype information to select which genotype to use as parents for crossing and select the progeny with an increased mean observation for the selected trait (Kingsolver *et al*., 2001). This is called ‘phenotypic’ selection (PS). With the coming age of marker-assist-selection (MAS), QTLs are increasingly used to provide prediction on the performance of phenotype based on the genetic information.

To perform selection based on genetic information, the first step is to identify the QTL and its effect on a given trait. QTL mapping uses maximum likelihood (ML) estimations to find intervals in the genome associated with the segregation of the phenotype observations. Conventional QTL mapping analysis is based on statistical analysis such as linear regression which allows for more computationally efficient predictions than ML (Kearsey & Hyne, 1994; Haley & Knott, 1992). Several software tools are widely used to perform QTL analysis including R/qtl and MapQTL (Broman *et al,* 2003; Oojien, 2004).

Using a regression algorithm for QTL analysis is computationally easier than the maximum likelihood method. However, regression methods require extensive amount of computational time to process a large marker dataset. To overcome this limitation, steps to simplify and reduce the number of markers must be performed. This will be implemented prior to analysis using simpler QTL methods, such as interval mapping in order to detect regions of interest or performed continually during analysis. When the low interest markers are identified, those markers can be removed to reduce the amounts of computation needed.

Once QTLs have been detected and markers in linkage have been identified, the next step is to apply this information to inform breeding. This approach is known as marker-assisted selection, or MAS. It is a method through which markers are used to predict the potential gain seen in progeny by selecting the suitable parents to increase/decrease occurrence of a particular trait. The aim of MAS is to outperform the PS by using genetic information and has been applied in many crop species (Prasanna *et al*., 2010; Steele *et al*. 2013; Ashraf & Foolad, 2013). However, this comes with an increased cost in both QTL detection and marker generation. Phenotypic selection is relatively straightforward with highly heritable traits as phenotypes are often simple to observe and therefore can achieve the easy gain in breeding. With genotyping costs falling dramatically for the last decade, using MAS in breeding becomes much more appealing. Also, MAS tends to be more effective in trait discovery where heritability is low (Van Berloo & Stam, 1998). Phenotyping is a time consuming and labour intensive process; hence being able to identify QTLs cost effectively could lead to a reduction in the reliance on costly PS to guide breeding programmes.

3.2.2 Conventional QTL analysis approach and bottlenecks

Conventional QTL analysis involves several processes. First, a mapping population has to be created and the suitable parents need to be selected to create the mapping population. Second, hundreds of progeny are created, planted, genotyped and observed. Higher numbers of progeny are preferred (Vales *et al.,* 2005). To create a genetic map, software is used for mapping including JoinMap, R/qtl and QTL cartographer (Broman *et al.,* 2003; Van Oojen, 2011; Wang *et al.,* 2012). Finally, QTL analysis is performed based on the genetic map.

With the coming of age of high-throughput genomic sequencing, large marker data sets are being utilised for QTL analysis. However, traditional software tools were not designed to effectively process high number of markers due to the extensive computation times needed for QTL analysis. Subsets of data are normally created based upon assumptions from other analysis to overcome the problem. Unfortunately, this will add bias as each iteration is only dealing with a subset of the data and risks the chance of QTL being missed due to bad selection criteria.

One general criticism of QTL mapping is that in practice only low numbers of potential QTLs are detected in many studies (Hyne & Kearsey, 1995; Kearsey & Farquhar, 1998; Laurie *et al*., 2004). It is now a consensus that traits are unlikely to be a combination of many large QTLs and are much more likely to be a mixture of a small number of large effect QTL plus many small effect QTLs (Buckler *et al*., 2009). One published study suggests that up to 50 QTLs could potentially be detected in one species (Laurie *et al*., 2004), however, this still may not cover all the QTLs controlling one particular trait. Several QTL mapping families may therefore be needed in order to identify all the QTLs involved (Jannink *et al.,* 2001; Rosyara *et al.,* 2009).

The potential gain by using conventional statistics-based QTL analysis approach to perform MAS may therefore be limited due to the high cost in manpower and lengthy computation time needed. A more efficient approach for QTL analysis is therefore needed to reduce the time and cost of MAS.

3.3 Random Forest for QTL Analysis

Conventional QTL analysis is based on statistical methods and as all methods comes with a set of biases and assumptions, such as the assumption that all observations are effecting the result. When the analysis is performed on high dimensional data, these assumptions tend to fail, and the methods begin to suffer from a long computation time. In order to tackle the bottlenecks that exist in conventional QTL analysis, a machine learning based analysis tool was developed. This tool simplifies the process of detecting QTLs with high sensitivity and has the ability to handle massive and complex datasets with reduced computation time.

Random forest (Brieman 2001a), based on one of the machine learning algorithms that uses both bagging and bootstrapping to develop a collection of decision trees known as a forest, was adopted to develop this tool. It can be used for both classification and regression problems and therefore can handle both categorical and numerical phenotypes.

Random forest builds each tree from a different subset of attributes, the number of which is defined by an mtry parameter. Each tree analyses a different problem space. The theory is, by repeating the same procedure over many trees, that a more general model of the problem will be detected by averaging the response of all the trees created. This process means that single noisy attributes can only influence a subset of the trees and therefore will not affect the whole model. It also allows the model better tolerance to errors

during genotyping.

Previous studies have used random forest to identify metabolites (Scott *et al*., 2010), mapping DNA aptamer fitness (Knight *et al*., 2009) and genomic selection (Heslot e*t al.*, 2012). All of these studies handled high dimensional data where the number of attributes far outweighs the number of observations. And these studies have demonstrated that random forest does have great potential in dealing with complex datasets such as those generated from high-throughput technology.

Random forest is an extension of decision trees which perform attribute selection by calculating the ability of each class for a given attribute to improve the understanding of the tree in a given problem. This suggests that only the most informative attribute is selected and those having lesser effect are not included. Each branch selected in the tree will change the remaining distribution of data. Selections are made on each branch point depending upon the branching points that occurred before. This implies that each selection is based upon the previous ones and relationships between markers are taken into account during the analysis.

Upon completion, the forest attributes can then be ranked using the importance metric. Importance is calculated as the average difference between the out-of-bag error before and after permutation across the forest.

These marker calls are then compiled into a matrix with p (number of markers, plus one for the phenotype observation) being the column and N (number of observations) as rows. Each row is the allele calls for that genotype along with the observed phenotype. An example of a matrix can be seen in Table 3.1. In this instance there are three (N = 3) genotypes with a single observation taken. This data set consists of 8 genetic markers with the phenotypic observation in the final column (p = 8 + 1), where calls are coded as A, B,

H or 0.

	M1	M2	M3	M4	M5	M6	M7	M8	Obs
Geno1	A	A	B	H	H	H	0	A	150
Geno2	B	A	B	B	H	H	A	H	175
Geno3	B	B	B	H	H	A	0	H	165

Table 3.1: Example data matrix used to train the random forest model. Rows consist of marker calls for each genotype. Columns represent markers, and rows are the observation.

Additional attributes for cofactors such as year of observation, plant age, or any other variables including environmental data can be added into the matrix when necessary.

Software tool R with randomForest library is used in this study to implement the random forest algorithms (Liaw & Wiener, 2002) and parallel computation is also employed to reduce the computation time through implementation of the R library SPRINT (Hill *et al*., 2008).

The importance scores were extracted from the resulting model and aligned with the genetic map to identify the regions where possible QTLs may exist.

3.4 Results and Discussion

A flowering time mapping family, Mx2, was established to study the genetic variation of flowering time in *Miscanthus*. The parents selected show variation in flowering time, with one early flowering and the other late flowering. The 185 progeny were selected and planted in three replicates at the same site.

The trial located at IBERS (Aberystwyth, Wales) was monitored for several years. However due to a harsh frost experienced in early 2009, several plants were lost and replanted with a clone in the same year. This replanting meant that some plants were younger than the others in the same trial. It is known that *Miscanthus* requires several

years to reach maturity so the age of the plant may have an effect on the performance of phenotype and therefore this discrepancy was put into consideration during analysis.

To test the capability of using random forest algorithm to detect QTL, the newly developed tool referred to as RFQTL, has been applied on the 3475 SNP markers generated from the Mx2 mapping population via genotyping-by-sequencing (Elshire *et al*., 2011). Markers were assigned allele calls for each genotype. They were coded as A, B for the homozygote and H for the heterozygote with an additional coding of 0 for unsuccessful allele call due to no or low sequence read coverage. A genetic map was created using this same dataset (Ma *et al.,* 2012b).

Flowering time information was collected over several years (Table 3.2). In 2013, a different flowering measurement was used due to the change of phenotyping strategy. The new phenotyping of flowering time was simplified from 4 stages to only 1 stage. Only flowering stage 2, panicle emergence, was measured but with an increased frequency in order to increase the number of data points and obtain a better quality flowering distribution.

Therefore the 2013 dataset were analysed independently from 2009-2011 analysis. The analysis of 2009-2011 was carried out using MapQTL, GoldenHelix and random forest and the results are presented and discussed in 3.4.1. Section 3.4.2 describes the results from QTL analysis of 2013 and the discussion on comparison between analysis methods. Measurements of stage 1 ‘flag leaf emergence’ and stage 2 ‘panicle emergence’ will be used to associate with markers in 2009-2011 analysis. As only stage 2 ‘panicle emergence’ was used for such associations.

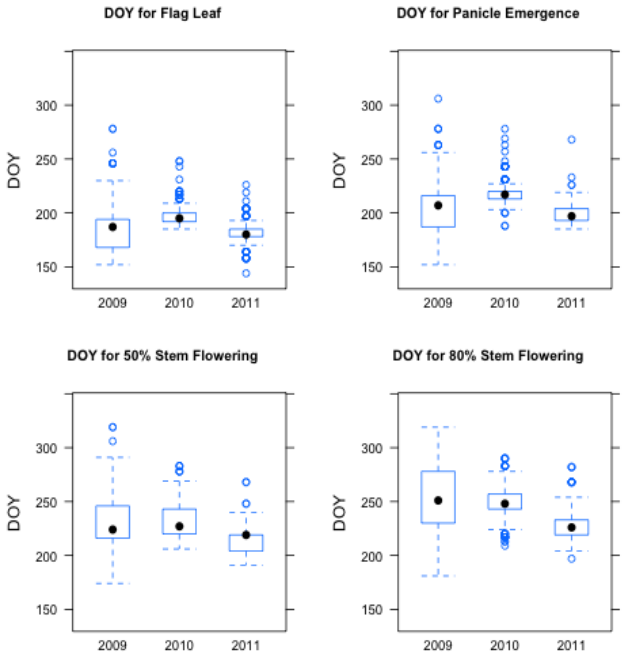

Figure 3.1: Flowering phenotype observations of the Mx2 mapping family. Each year has a different distribution possibly caused by meteorological differences between the years.

Years	Observation	Observation Frequency	Methods Applied
2009-2011	Flag Leaf Emergence (Flowering Stage 1)	Weekly	RFQTL, MapQTL, GoldenHelix
2009-2011	Panicle Emergence (Flowering Stage 2).	Weekly	RFQTL, MapQTL, GoldenHelix
2013	Panicle Emergence (Flowering Stage 2),	Twice per week	RFQTL

Table 3.2: QTL analysis on Mx2 flowering time mapping population

3.4.1 2009 – 2011 Flowering Time QTL Analysis

Over the three year period between 2009 and 2011, flowering time was scored using a four stages system outlined as in Chapter 2.4.3. The four stages are flag leaf emergence, panicle emergence, 50% stem flowering and 80% stem flowering. These measurements were either taken once a week or in a few instances once every fortnight. There are variations in the distribution of flowering between each year, caused potentially

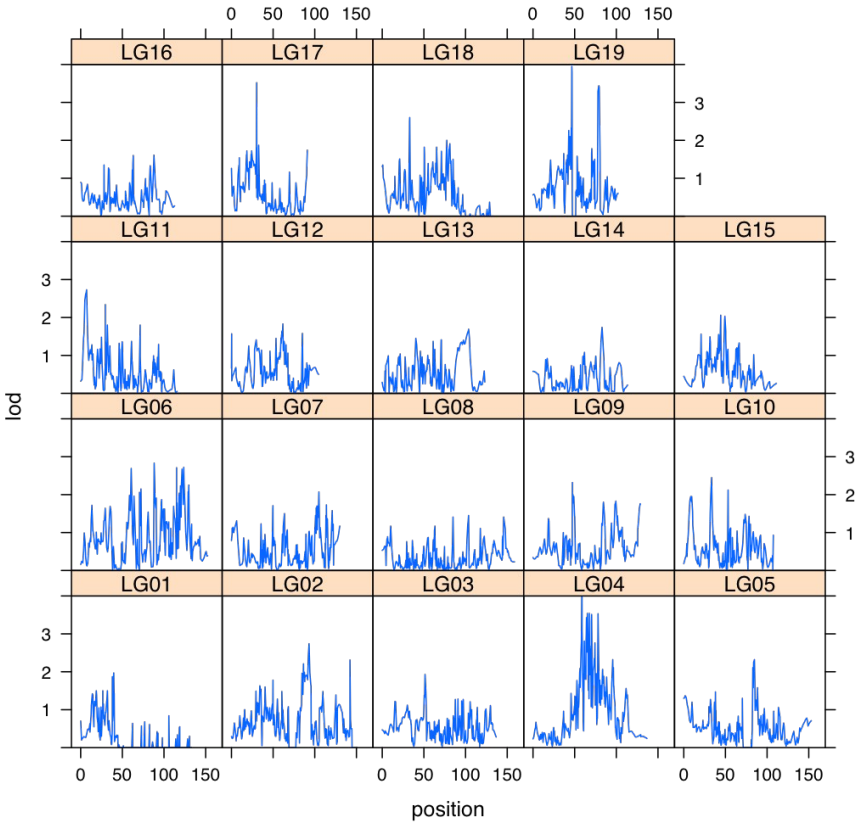

Figure 3.2: Interval mapping was applied on flag leaf emergence data. This is the simplest and the fastest approach for implementing QTL mapping with a potentiality QTL on LG04, 17 and 19.

by environmental differences between years (Figure 3.1). Jensen *et al* (2011a) found that flowering time fluctuates significantly between years in both *M. sinensis* and *M. sacchariflorus*. *M. sinensis* has been suggested to be day length neutral (Deuter, 2000). Another study of *Miscanthus* flowering time has also shown drought stress might delay the flowering time in *Miscanthus* (Jensen *et al*., 2011b).

All analysis only used the first two stages of flowering time measurement as the last two stages were flowering intensity. The phenotype data from stage 1 of flag leaf emergence and stage 2 of and panicle emergence were used for QTL analysis using MapQTL and GoldenHelix and random forest approach RFQTL. Comparison between 3 results were performed for cross validation. Environmental factors were taken into consideration to refine the model.

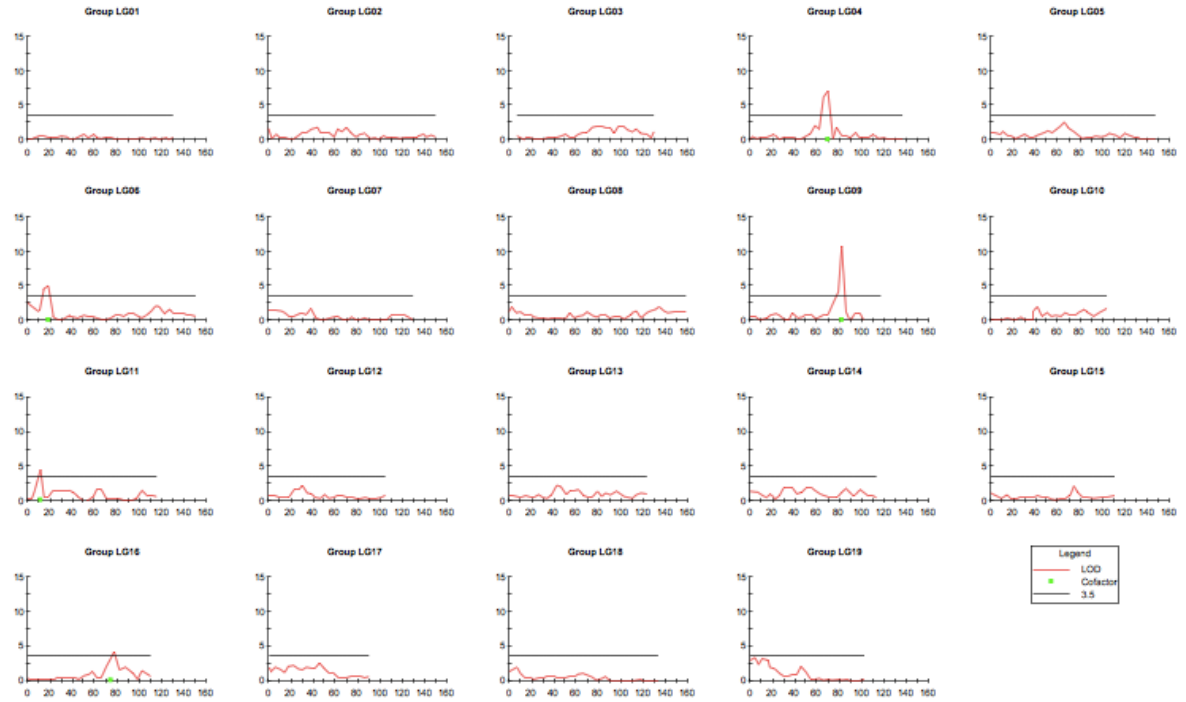

Figure 3.3: MapQTL results from MQM mapping of flag leaf emergence.QTLs appear to exist on LG04, LG06 and LG09, LG11 and LG16.

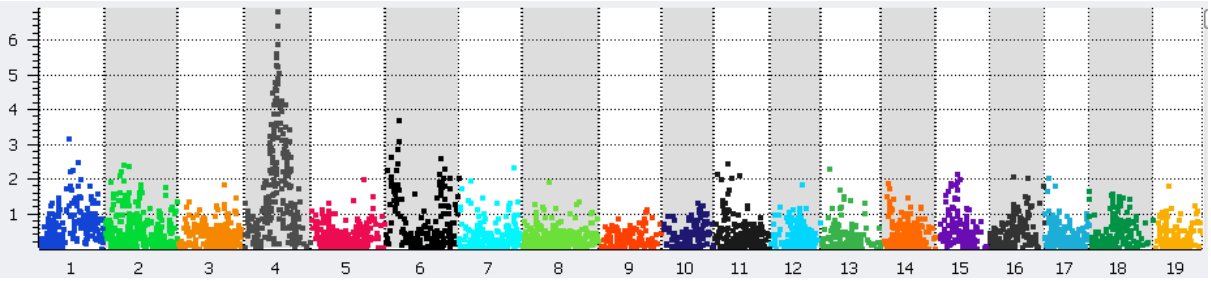

Figure 3.4: The GoldenHelix results of flag leaf emergence. Markers that relate to flag leaf emergence appear on LG4 and LG6.

Stage 1 Flag Leaf Emergence – MapQTL and GoldenHelix

Flowering stage 1 is reached when the plant produces a flag leaf. This is the last leaf a stem will produce and it will be followed by the emergence of a panicle. This leaf is much thinner than normal leaves and is the first indication that the plant is transitioning from vegetative growth into reproductive growth.

Interval mapping was performed using MapQTL for the flag leaf emergence data. There is strong indication of a potential QTL for flag leaf emergence located on LG04 with

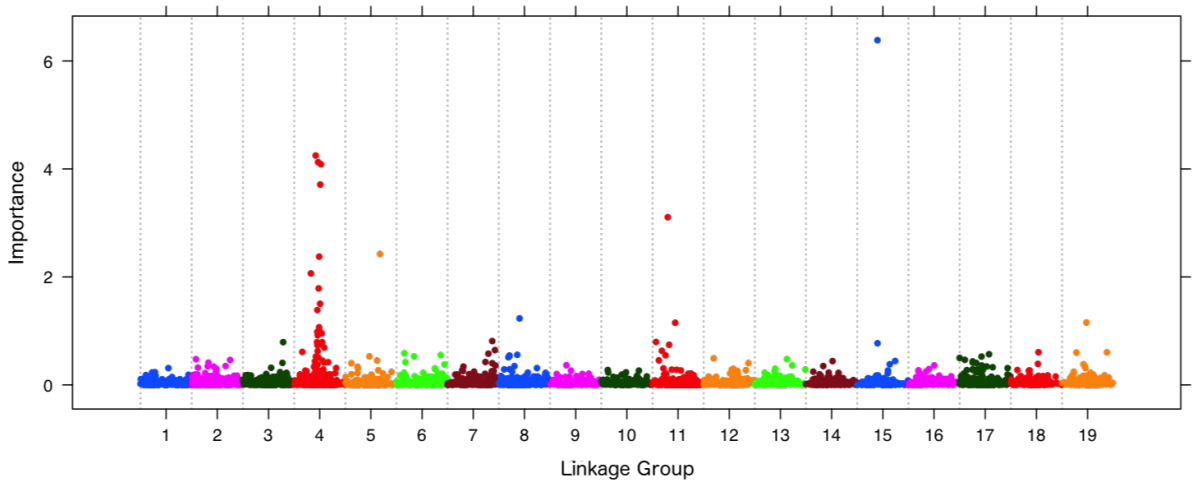

Figure 3.5: The importance scores from random forest QTL analysis RFQTL for flag leaf emergence plotted onto genetic map position. There is a major grouping of significant markers clustering on LG04, with one single important marker appears in LG5, 11 and 15.

possible linkage on LG19 and LG17 (Figure 3.2). It is a faster approach, but it is unlikely to correctly identify exact locations of QTLs. This method often just highlights the regions where potential QTLs might be located.

Both MapQTL (using the MQM method) and GoldenHelix have been deployed to conduct QTL analysis using SNP markers generated from this mapping family (Donnison *et al*, in preparation).

We have observed strong QTL signals on LG04, LG09, and several others on LG06, LG11 and LG16 that just reach the LOD level from the MapQTL results (Figure 3.3). Many of these peaks are consistent with the results from GoldenHelix (Figure 3.4). However there is no sign of any marker trait association occur on LG09 in GoldenHelix's result.

Stage 1 Flag Leaf Emergence – RFQTL mapping

Random forest based QTL analysis, RFQTL, was applied to analyse the same dataset (Figure 3.5). Two additional factors were included in this analysis, the year and age, to take into account the different ages of the plants and multi-year observations. The

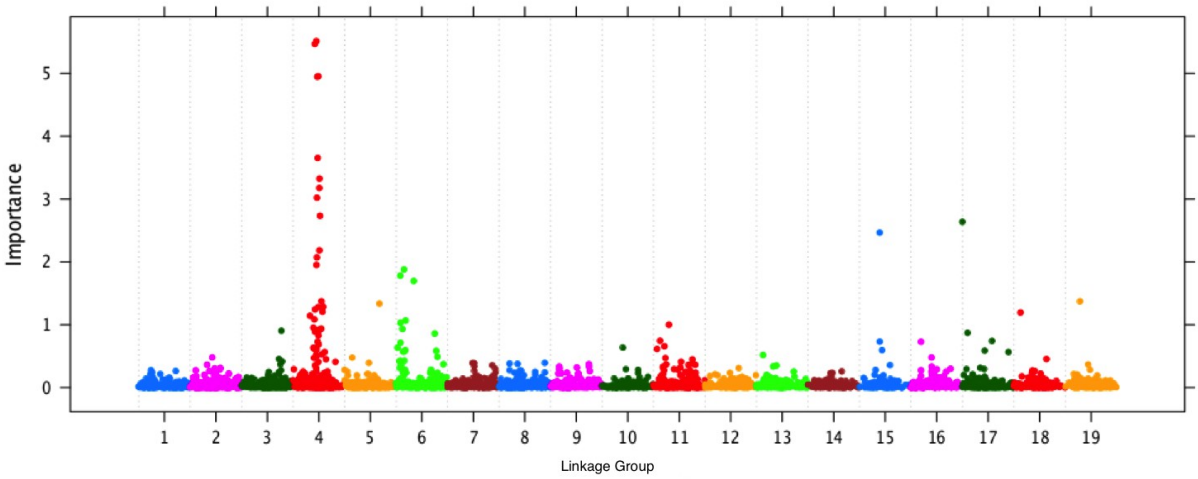

Figure 3.6: The results from RFQTL including the meteorological data by taking into account for variances in years and age. There is a strong signal in LG04 again, but the marker with high importance score appeared previously in LG15 has now decreased its importance. A new significant region appears in LG06.

important scores of age is 28.7 and 51.77 for year. The study results have shown the environment where the plant were grown and the age of a plant are both highly significant factors for determining the flowering time.

The top markers from RFQTL results lie between 58cM to 72cM in linkage group 4 with one single marker appearing in linkage group 5,11 and 15. Age and year are seen to have a large effect in this model. This implies that the environmental factors have a considerable influence on flowering time in *Miscanthus*.

Stage 1 Flag Leaf Emergence – RFQTL model with environmental variances

To explain the environmental variances between years and to refine the model with higher sensitivity, the meteorological data was included in the analysis to allow the random forest to detect the differences between each year.

The meteorological data was parameterised using rainfall, PAR (photo-synthetically active radiation), minimum and maximum temperature. Meteorological data came from the Met Office station at the Gogerddan site of IBERS located less than a mile from the trial

site. The meteorological observations were parameterised using the method outlined in chapter 2 and this yielded 37 attributes for each observation. A total of 148 meteorological observations were added into the model.

RFQTL Mapping with Meteorological Observations

RFQTL was performed on the data set with the additional meteorological attributes (Figure 3.6). The year and age had important score of 0.15 and 1.1, respectively. The year is insignificant when compared with the top marker which usually has the score over 5. Age appears to have some effects on the model but is now less significant. This result suggest that maturity does have effect on flowering time in *Miscanthus*.

Stage 1 Flag leaf Emergence – Method Comparisons

To compare the results between RFQTL, (Figure 3.6), with GoldenHelix (Figure 3.3) and MapQTL (Figure 3.4), it is obvious that all three methods highlight the same regions of importance on LG04 and LG06. Also, results from all three methods appeared to suggest that the beginning of LG11 is a potential region containing smaller effect markers. RFQTL method did identify potential significant markers on LG15 and LG17. However, those markers did not appear in the results from GoldenHelix and MapQTL. These markers were also missing in the RFQTL results without including meteorological attributes. This could imply that without meteorological data the RFQTL method misses many potential QTL. This could be due to the fact that the RFQTL method can use any attribute to explain variations. Markers can be used to explain variances that were environmental in nature, therefore missing actual QTL. Possible explanations for single markers with high importance could be that it is a false positive or that a potential marker was mapped to the wrong location since random forest does not use position information in its analysis. On the other hand, conventional analysis makes use of the position information in the

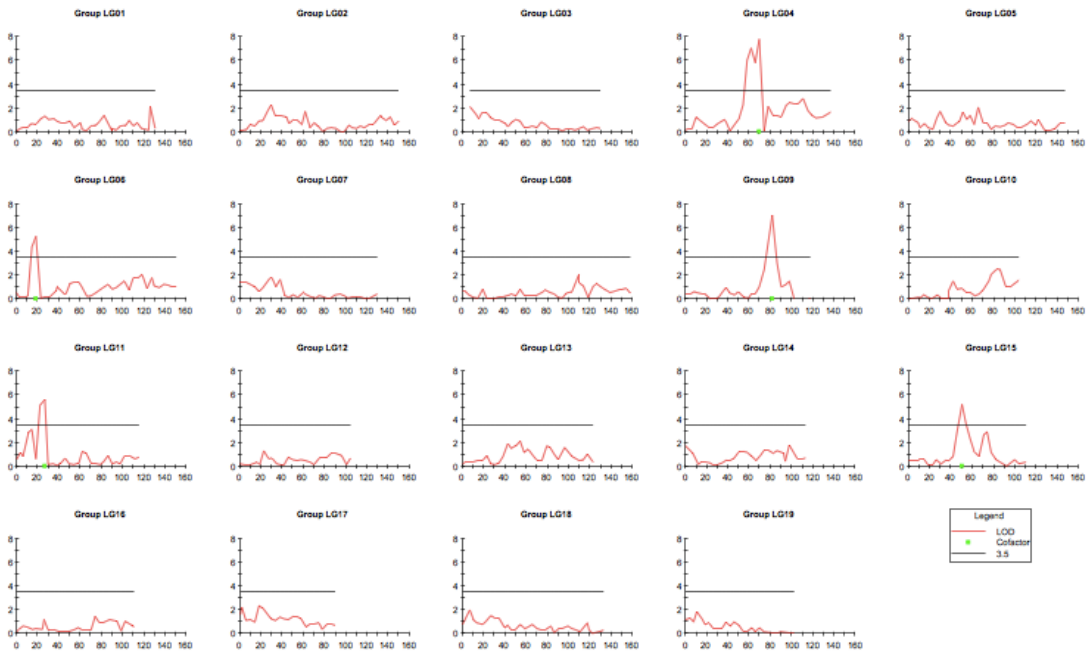

Figure 3.7: The analysis results from the MapQTL analysis on panicle emergence data. QTLs appear on LG04, LG06, LG09, LG11 and LG15.

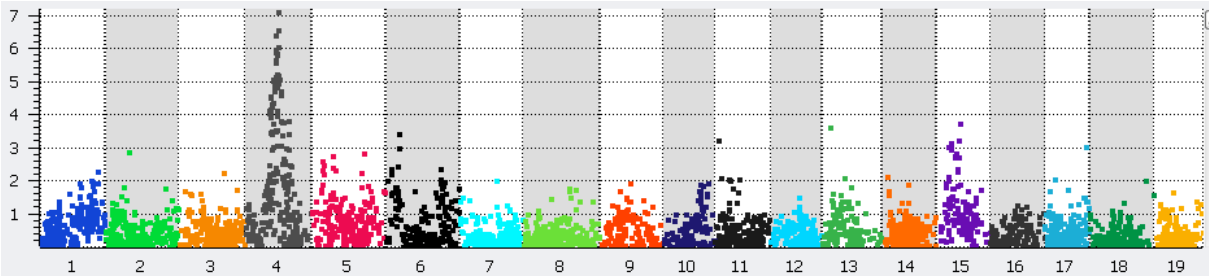

Figure 3.8: The analysis results from GoldenHelix analysis of panicle emergence. Regions show high marker relationship on LG4, 6, and 15. Single markers appeared LG 11 and 13.

analysis. It attempts to detect QTL between markers, using the genetic map to order them. RFQTL instead investigates markers which explain variance regardless of position therefore markers will be detected without consideration to the neighbouring markers. This may give RFQTL the potential to detect associations that would be missed due to incorrect mapping. Further investigation is necessary to find out the cause of this disparity. MapQTL results suggested that there is a potential QTL on LG09 but neither GoldenHelix or RFQTL detected any relationship between flag leaf emergence in this region and this discrepancy

should be due to multiple-QTL model used in the MapQTL mapping. In conclusion, it would appear that GoldenHelix and RFQTL generate similar results but do not concur with some of the results from MapQTL.

Stage 2 Panicle Emergence – MapQTL and GoldenHelix

The emergence of a visible panicle is the second stage in *Miscanthus* flowering. This is scored when the panicle is approximately one centimetre visible from the leaf. Analysis of the panicle emergence data have been performed using MapQTL (Figure 3.7) and GoldenHelix (Figure 3.8) methods (Donnison *et al, in preparation*).

Stage 2 Panicle Emergence – RFQTL

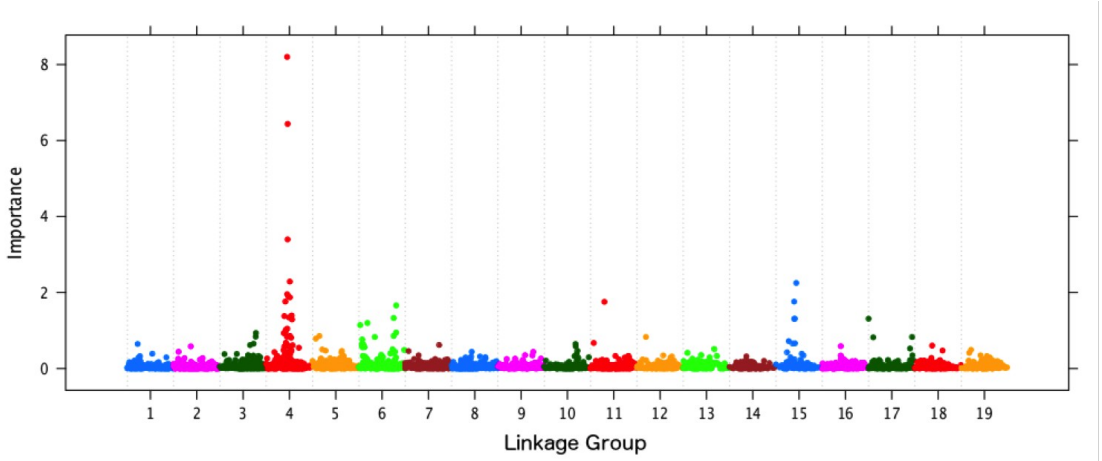

Figure 3.9: Importance scores from the RFQTL analysis of panicle emergence. This result shows potential QTLs on linkage group 4, 15 and possibly on 6 and 11.

Random forest QTL analysis, RFQTL was used to analyse the panicle data set (Figure 3.9). As discussed in previous section, meteorological data can improve the performance of the RFQTL to detect more QTL loci. Therefore meteorological data was also included when analysing panicle emergence. Their importance scores ranged from 3.22 to 0.0075. It highlights the fact that the environmental factors have a significant effect on the phenotype. Again, the age and year were included and their importance scores

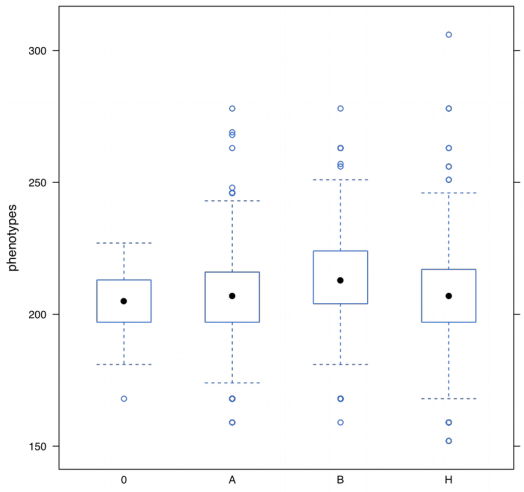

Figure 3.10: Distribution of top marker for panicle emergence which was selected by the RFQTL analysis.

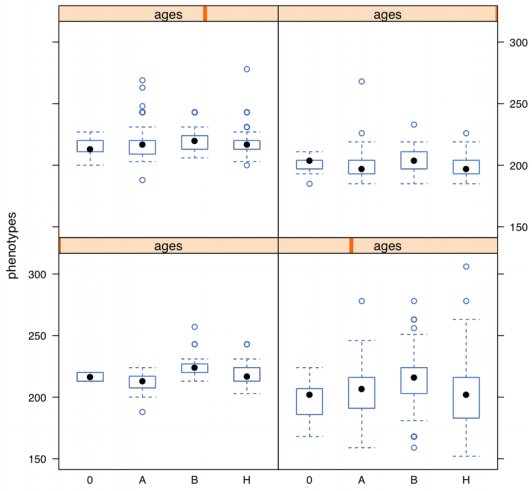

Figure 3.11: Distribution of the top marker for panicle emergence split into different groups based on age difference

were 1.33 and 0.31 respectively. This once again suggests that the age of a plant has strong influence on flowering time as the year has much smaller effect. The low importance score of year did suggest that some environmental variances might still be unaccounted for. The most significant SNP marker appears to locate on LG4 at 62cM with the second and third important SNP markers reside at 63cM and the forth most important SNP marker appears at 70cM.

Stage 2 Panicle Emergence – Method Comparisons between MapQTL, GoldenHelix and RFQTL

When comparing the results between MapQTL (Figure 3.7), GoldenHelix (Figure 3.8) and RFQTL (Figure 3.9), there is strong indication of significant QTL presented on LG04 in all three analysis results. They also all detected potential QTL on LG06 and LG15. MapQTL detected an extra QTL on LG09, but no association was found in either RFQTL or GoldenHelix results.

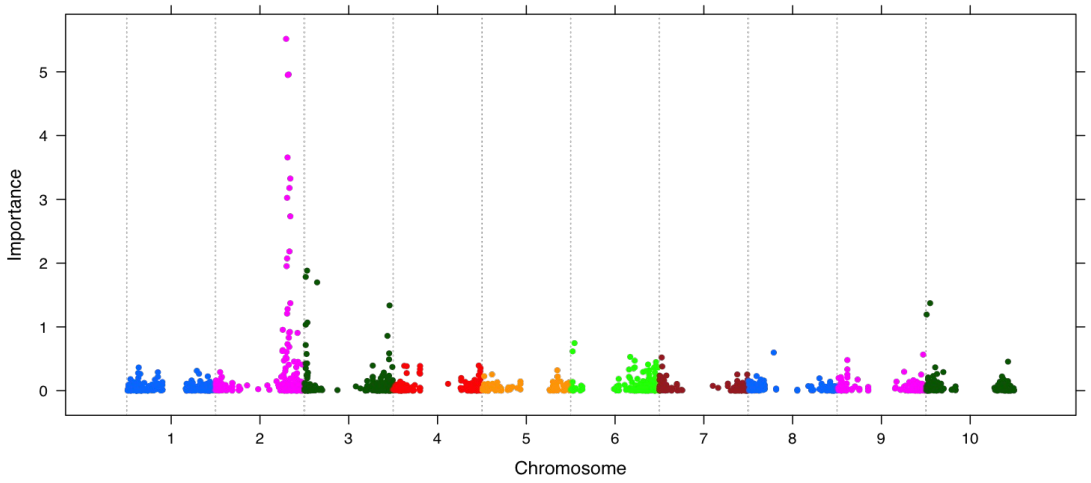

Figure 3.12: The importance scores from analysis of Micsanthus flag leaf emergence were mapped onto the Sorghum genome. The importance region responsible for flowering time highly align with Sorghum chromosome two.

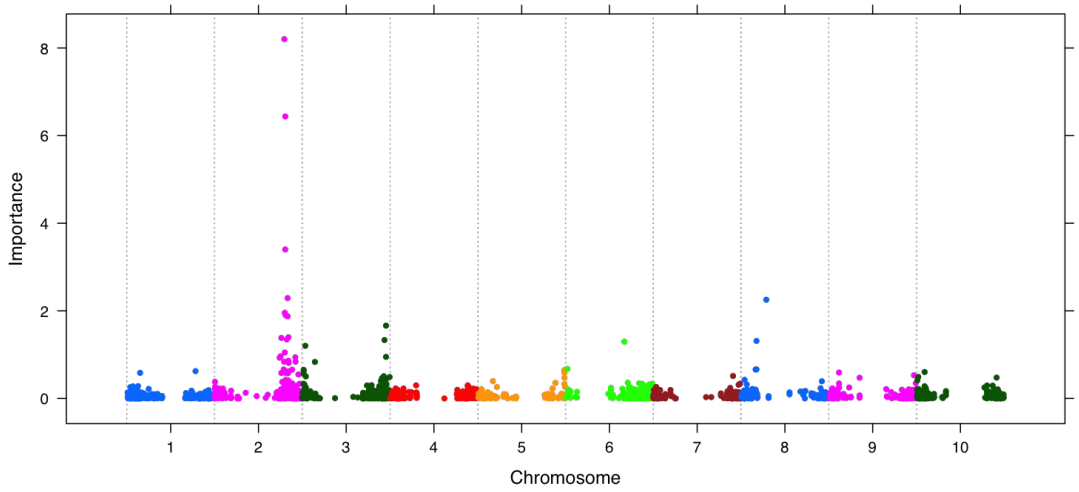

Figure 3.13: The importance scores from RFQTL of panicle emergence were mapped onto the Sorghum genome. A peak of high importance on chromosome 2 was observed with other high importance markers appearing on chromosome 8 and 3.

Stage 2 Panicle Emergence – High Importance Markers

The distribution of the top marker for panicle emergence has been plotted against the allele call for each genotype with the B allele shows association with later flowering (Figure 3.10). The same data is split into separate groups based on age difference. Nevertheless regardless of age, the B allele is always associated with the later flowering phenotype (Figure 3.11).

Stage 2 Panicle Emergence – Comparisons with Sorghum

There is clear evidence that the high synteny exists between *Miscanthus* and *Sorghum* (Swaminathan *et al.,* 2010). Hence, the *Miscanthus* markers used for MapQTL and GoldenHelix had been mapped onto *Sorghum* genome. The importance scores generated from the RFQTL were subsequently mapped onto the *Sorghum* genome for comparisons. It has been demonstrated that the important region in *Miscanthus* LG04 aligns with Chromosome 2 of *Sorghum* (Figure 3.12).

The most important marker located in Chromosome 2 lies within the region of a documented flowering QTL in *Sorghum*. This QTL is pointed out in the Comparative Saccharinae Genome Resource (CSGR)-QTL database, (Zhang *et al*., 2013). It is located in the region between 61,550,813 bp and 66,088,144 bp (Lin *et al*., 1995).

The importance scores from panicle emergence QTL mapping have been mapped onto the *Sorghum* genome (Figure 3.13). The region of high importance once again appears around the end of chromosome 2 of *Sorghum*. All the markers that were mapped tend to cluster around the end of Chromosome 2 with the most important one being located at 61,861,092 bp. It also observed a peak on Chromosome 8 at 15,912,450 bp and Chromosome 3 at 70,978,410 bp.

The highest importance score located on Chromosome 2 of *Sorghum* is the same region as found in the flag leaf emergence analysis results. Again this region also lies within the known QTL responsible for flowering time in *Sorghum* (Lin *et al*., 1995) as previously discussed in the stage 1 flag leaf emergence analysis results.

The other two peaks appear in Chromosome 8 and Chromosome 3 do not match any QTL found in the (CSGR)-QTL database. Further investigation would be required to

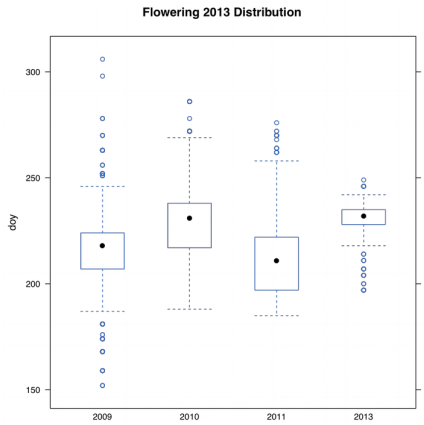

Figure 3.14: The distribution of flowering stage two (panicle emergence) for year 2009, 2010, 2011 and 2013. We saw a late average flowering in 2013 similar to that of 2010, but the range was much decreased much more.

confirm their association with flowering time in *Miscanthus*. It could be either these are yet to be identified in *Sorghum* or are newly found QTLs controlling flowering time in *Miscanthus*, or false positives.

3.4.2 2013 Flowering Time QTL Analysis

No flowering measurements were taken in 2012. In 2013 flowering measurements were taken at higher frequency but only flowering stage 2 and panicle emergence were measured to improve time efficiency in the phenotyping process. Each genotype was scored twice a week. This data was then processed to create a DOY value for panicle emergence.

Due to the change in phenotyping method in 2013, only random forest method, RFQTL, was applied to the 2013 dataset. MapQTL and GoldenHelix were not used to analyse 2013 dataset so there are no results from these methods for comparison.

The 2013 observations were plotted against the measurements taken from the 2009-2011 flowering stage 2 panicle emergence (Figure 3.14). The flowering in 2013 seemed to occur later than seen in 2009 and 2012. They are much more in line with the 2010 single

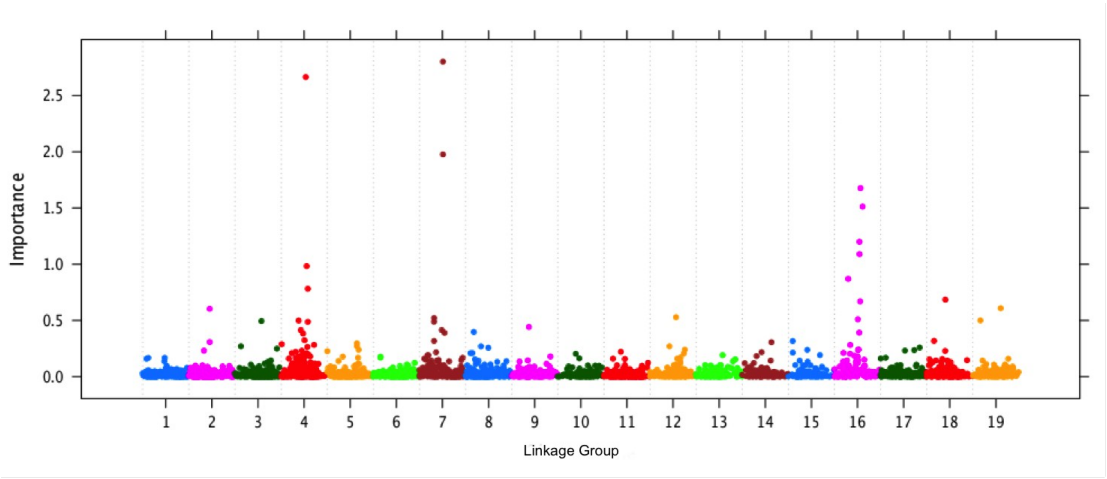

Figure 3.15: Importance score of 2013 panicle emergence from RFQTL analysis. There are similar high important markers on linkage group 4 but high importance markers also appear on linkage group 7 and 16.

year observations. 2013 was a year of high temperatures with low rainfall during the flowering period. This finding is consistent with the study by Jensen *et al* (2011b) which concluded that the flowering time can be delayed due to drought.

Observations were taken only within one single year and no meteorological attributes were included in the analysis, as all plants were assumed to have experienced the same environmental conditions. Age was still included. The aim of this experiment was to learn if an increase in phenotyping frequency, but only performed over one year, would produce the similar QTL results as found in the 2009-2011 analysis. Therefore only stage 2 panicle emergency measurement was used for this 2013 analysis.

2013 Panicle Emergence – RFQTL

The importance scores from the 2013 observations have been shown on the genetic map (Figure 3.15). The age attribute had an importance of 0.99, slightly less than in previous studies.

Two new QTLs emerged, one located on LG07 and another located on LG16. The QTLs seen on LG04 in previous 2009-2011 studies were once again detected on LG04.

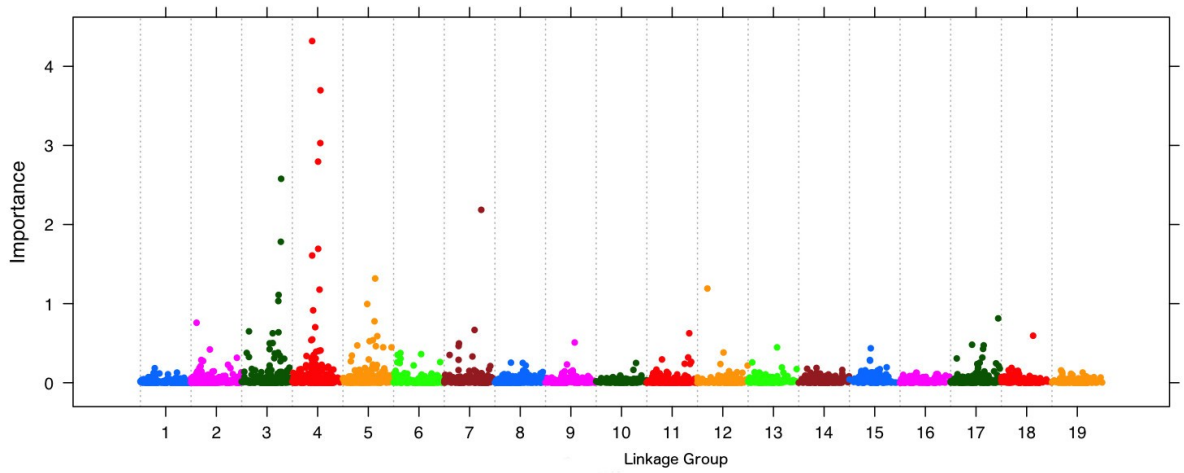

Figure 3.16: Importance score of 2010 panicle emergence from RFQTL analysis. There is high importance seen on LG04 and LG03. An important single marker appears on LG07.

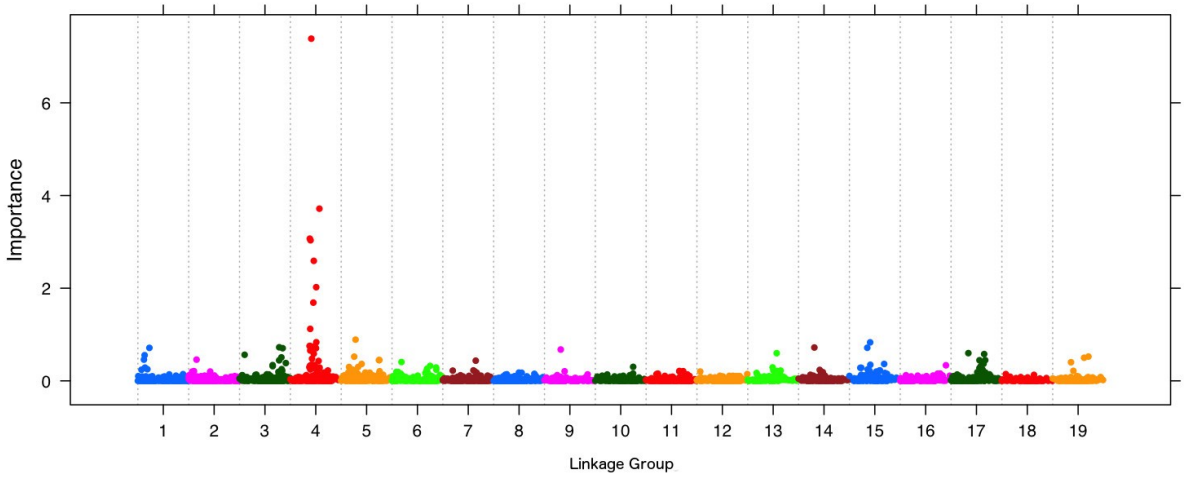

Figure 3.17: Importance score of 2011 panicle emergence from RFQTL analysis. There is a high importance peak seen on LG04. Potentially there is a number of small effect QTLs across the genome.

For comparison, each year from the 2009-2011 studies were modelled using RFQTL. The RFQTL was unable to explain the variance in the 2009 study. However, it was able to explain the variances in the 2010 dataset (Figure 3.16) and 2011 dataset (Figure 3.17). In the three separate year experiments, the QTL on LG04 is always present. There was a high importance marker detected on LG07 in the 2010 dataset (Figure 3.14). The 2010 flowering which is similar to that seen in 2013 was later than in 2011. However the potential QTL on LG16 was only found in the 2013 results for panicle emergence analysis.

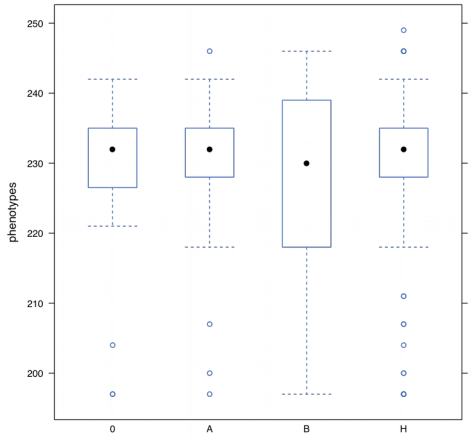

Figure 3.18: Top marker allele versus phenotype distribution of panicle emergence in 2013.

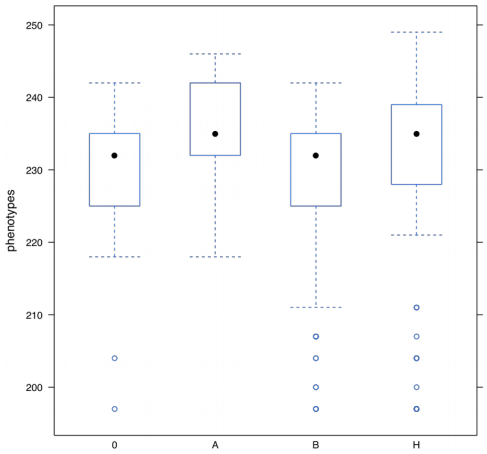

Figure 3.19: Second most important marker allele versus phenotype distribution of panicle emergence in 2013.

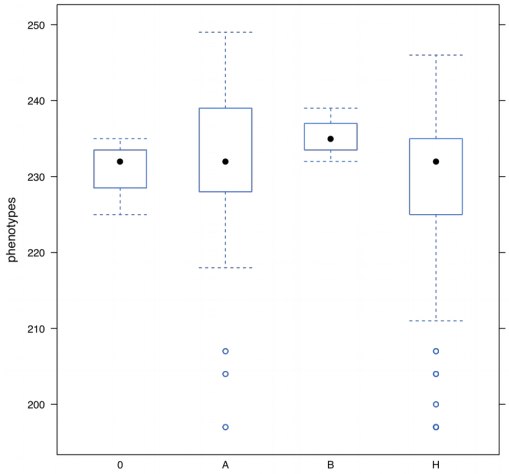

Figure 3.20: The most important marker appearing on LG16 of panicle emergence in 2013.

A QTL was detected in the same region in the flag leaf emergence analysis using MapQTL.

2013 Panicle Emergence – High Importance Markers

The phenotypic observations of 2013 were plotted against the alleles of the most important marker (Figure 3.18). The marker revealed that the B allele is highly associated with the variation of DOY. The phenotypic observation versus the alleles of the top marker in LG04 is shown (Figure 3.19). The B allele had shown strong association with the early

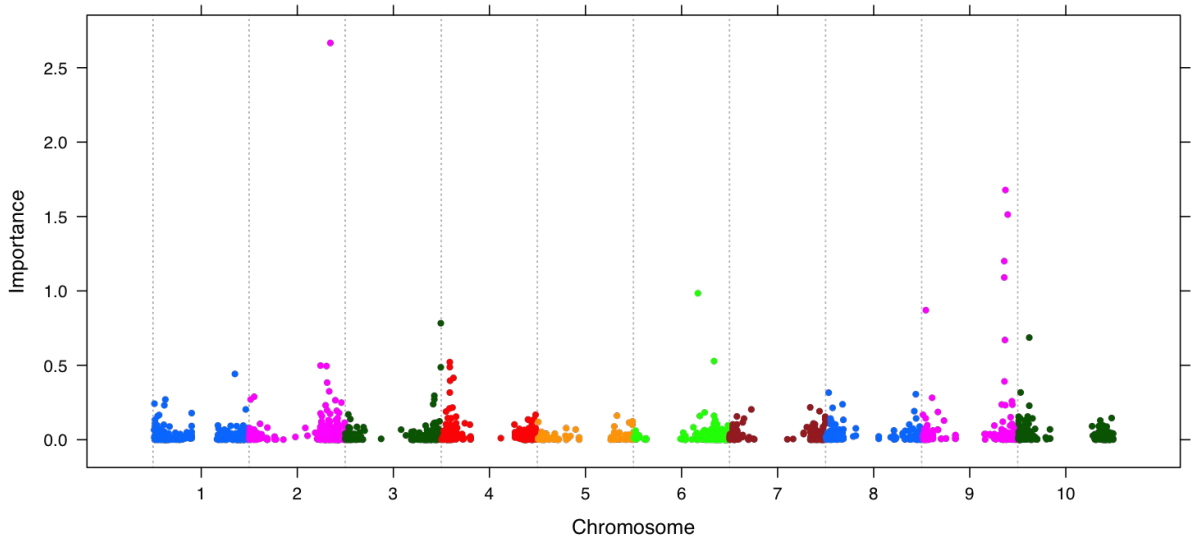

Figure 3.21: The importance scores of panicle emergence from 2013 analysis plotted onto the Sorghum genome. There are similar high importance markers on chromosome 2, but new cluster on chromosome 9 appeared. One high importance marker seen in linkage group 7 couldn't be mapped to the Sorghum genome.

flowering phenotype. The top marker on LG16 indicated that the B allele appears to associate with the late flowering phenotype (Figure 3.20).

2013 Panicle Emergence – Comparisons with Sorghum

The importance scores of the 2013 analysis were plotted against the *Sorghum* genome as shown in Figure 3.21. An interesting finding from this exercise was that the top importance marker from LG07 was not able to be mapped to *Sorghum* genome. Although the marker mapped to Chromosome 2 is still within the region previously presented in 3.4.1. The top marker mapped on to Chromosome 9 of *Sorghum* is located at the position of 51,943,327bp, which lies within a known QTL described in (CSGR)-QTL database located between 8,143,590bp and 57,010,750bp (Lin *et al*., 1995). The high importance marker seen in Chromosome 6 at 41,553,290bp lie within another documented QTL located between 38,005,075bp and 45,215,973bp (Lin *et al*., 1995).

3.4.3 Discussion

QTL detection

By comparing the analysis results from three methods on the studies of flowering time mapping population Mx2, there is a clear consensus that a major flowering time QTL is located on LG04 (Figure 3.3, Figure 3.4, Figure 3.6, Figure 3.7, Figure 3.8, Figure 3.9, Figure 3.15, Figure 3.16, Figure 3.17). The effects of the QTL controlling this trait become obvious when the markers selected by RFQTL were plotted against the phenotype (Figure 3.10, Figure 3.19). It would appear that this particular QTL alters flowering time by several days.

The region detected by the MapQTL analysis spans a wider range of cM, and within that range there appears to be two peaks. The RFQTL also detects these two peaks, one between 62.3cM and 62.54cM, and another one at 70cM. The MapQTL placed the QTL at the 70cM loci. GoldenHelix shows a large peak that spans a wide section of the linkage group 4 in the same loci. RFQTL results shows a much sharper peak, with a high importance at the 62cM loci. The surrounding markers are all seen to have less importance. This could imply that importance is related to linkage disequilibrium (LD) that the QTL is located near to 62cM. Further study using simulations is required to investigate the relationship between LD and importance scores. The RFQTL did also suggest high importance markers at the 70cM locus, which is the same locus found by MapQTL analysis. However this locus was not the most important in the RFQTL model. The markers in this position was ranked at the 5th and 10th.

When performing the 2010, 2011 and 2013 single year analyses, the peak on LG04 was detected in different locations. For 2010 and 2011, it appeared to be at 56cM and

53cM respectively. However in both instances many of the high importance markers were found around 70cM. In the 2013 analysis results the marker was detected at 72cM. This could imply that the higher frequency of observation used in 2013 allowed for more accurate QTL locus detection. Although it would appear that analysis of multiple years observations give better ability to detect multiple QTL when using RFQTL.

The QTL detected on LG09 was only found in MapQTL results as RFQTL and GoldenHelix do not show any marker effects in LG09. This indicated the potential for false QTL being detected by either MapQTL or by the RFQTL and GoldenHelix methods. Future study would be needed to confirm or reject the existence of QTL found on LG09.

The 2013 analysis did however reveal several new potential QTLs that may relate to the response to heat and possible drought stress. The region is known to contain a flowering QTL but was only detected in a year with high temperature and low rainfall. The potential QTL on LG07 does not share a strong match with the models of dominance for alleles, as seen in Section 2.2.1. Instead it appears that several of the alleles results in a more regulated flowering, whereas the B allele relates to a much wider variation (Figure 3.18). Further study of QTL would be needed to understand what processes lead to the observed distribution.

Comparisons have been made on the major QTL discovered. Under closer inspection of the results from RFQTL, many more markers with high importance across the genome were identified (Figure 3.9). Many of the markers detected seem to display some small association with the trait, due to them having an importance which was greater than zero. MapQTL excluded these regions that did not have a high enough LOD score to be considered as QTL. However as discussed in some literature, many traits are believed to be controlled by small effected QTL (Buckler *et al*., 2009). Further study of these low score

markers may reveal the missed QTL that were not detected by conventional QTL analysis methods.

Comparisons with Sorghum

The most important region identified by all 3 methods is located on linkage group 4. When they were aligned to the *Sorghum* genome, they were found near the common loci known to control the *Sorghum* flowering time (Lin *et al*., 1995).

It is widely recognised that *Miscanthus'* evolutionary origin was a divergence from the genome duplication of *Sorghum* (Swaminathan *et al.*, 2012; Ma *et al.*, 2012b). Although it appears that only one of the two copies of the chromosomes is in control of flowering time. It is well recognised that duplicated genes can be lost after genome duplications. But other interactions, such as epistatic silencing and differential expression, have also taken place after duplications (Adams & Wendel, 2005). Any of these hypotheses could explain why only one of the two chromosomes inherited from *Sorghum* still displays a potential QTL for flowering. Similar results were seen in another QTL study of a *M. sinensis* cross (Gifford *et al*. 2014). Gifford *et al* also detected a flowering QTL within a single linkage group. However it did not appear on the second linkage group that shared similar synteny with *Sorghum*. The authors also suggest the lack of a second QTL could be caused by gene loss or epistatic silencing.

Family Size

The number of progeny used in this study is comparatively small for QTL analysis. It is preferable that at least an extra one hundred progeny should be included to increase accuracy and QTL detection. The effects of population size are known to have profound influence on the analysis results from conventional QTL analysis methods. However, the

effect of family size toward the machine learning approach is still unknown and warrants further investigation. Increasing progeny number can increase the chance of cross over, so logically it should increase the capability to detect QTL using a random forest approach.

Observation frequency

In the 2013 study, flowering was observed at a higher frequency but for only one year. The analysis successfully detected a QTL in the same region found in the 2009-2011 studies with a slight shift in its position. More regions with potential QTLs were detected in the multi-year study than single year study.

Comparing the results from a single year with multi-year analysis, it was suggested that the multi year analysis has the potential to detect more QTLs. Nonetheless, it does require the inclusion of meteorological data to improve the quality of QTL analysis. As a result, we can conclude from this study that multiple year analysis, with higher frequency measurements will improve the quality of data and capability of QTL detection.

Environmental effects on flowering time

There is strong evidence that *Miscanthus* flowering is highly affected by various environmental factors in different year. The inclusion of the meteorological data as attributes reduced the importance of the year attribute, but did not completely rule out the year attribute from the model. It could be that the parameterisation process is losing some resolution due to the 10 day measurements interval or that the 4 attributes measured do not have the resolution needed to identify all the QTLs.

Study	Age Importance
2009 – 2011 FS1 (no meteorological data)	28.7
2009 – 2011 FS1 (meteorological data)	1.1
2009 – 2011 FS2 (meteorological data)	1.33
2013 FS2	0.99

Table 3.3: The summary of age importance scores from 4 different studies with various attributes using random forest. It is obvious that without meteorological data the age attribute is much larger. However when meteorological data is added to the model the age variable's importance is much smaller. It does however appear in every model as an important feature.

Age

One interesting discovery from using random forest to detect QTL is that age appears to be an important attribute for flowering time in *Miscanthus*. A summary of scores for age show that even when meteorological data was included age is still an important attribute (Table 3.3). We see the biggest difference in the importance of age between those experiments that included meteorological data and the one which did not. This could imply that the age variable was allowing the random forest some attribute to which it could assign some of the variance caused by the different weather experienced in each year. Only a slight difference in the importance was seen between experiments that included meteorological data. Even in the 2013 study, where the plants should have now reached maturity, the age is still affecting flowering time. However the importance attributed to the age attribute could represent some other effect. The age variable could be considered to be the representation of the year of planting. This would imply that the planting conditions are still having an effect on phenotype even in a mature plant. Another explanation is that the plants used to replace the frost killed plants may have been a better quality, or the genotypes which survived the harsh frost are still suffering the effect of frost. Although age is important, its effect may only be apparent in some of the QTLs. The phenotype of the major QTL detected on LG04 does not appear to be affected by age difference (Figure 3.11).

Strength of Random Forest

One major advantage of the random forest approach is the simplicity and speed through which results are achieved. It does not need any pre-processing required in conventional approaches such as MapQTL. The random forest method uses raw allele calls and is simpler to implement from raw data. The RFQTL was proven to be faster than MapQTL, with the RFQTL method requiring only one hour for data processing for this particular dataset whereas the MapQTL approach was unable to complete the analysis for the whole dataset in a reasonable time. Instead data reduction had to be applied for the data to be processed. As the data set size increases so does the computing time of the random forest, with the rice data analysis described in chapter 2 taking approximately 3 days to complete. However this scaling could simply be down to implementation, the sprint library utilised has to make multiple data copies due to its implementation, which is a time consuming process. This could be improved upon, and potentially reduce the computing time further.

The ability of random forest to make decisions based upon previous observations, due to the tree structures it generates, allows random forest to detect effects between markers which is much more computationally intensive to do using conventional analysis. This could potentially reveal new relationships that have been missed using the standard analysis approaches. This differs vastly from the standard approaches used in QTL mapping, as outline in chapter 2, as these only look for comparisons between neighbouring markers to look for probable QTL locations. It does not account for interactions between QTL throughout the genome. The MQM method does account for effects from having multiple QTL but still does not account for interactions, such as epistatic effects.

3.5 Concluding Remarks

It is the purpose of this chapter to demonstrate that random forest is a powerful tool for QTL analysis. The use of random forest and importance scores in this research has demonstrated that an algorithmic model can correctly identify attributes and rank their effect in a genomic context. Additionally, random forest has increasingly been used in genomic selection (GS) as an alternative approach to utilise genetics in breeding (Heslot *et al*., 2012).

Conventional QTL analysis involves complex data processing and pre-processing to allow the analysis to handle large and multi-dimensional datasets. Random forest on the other hand does not require pre-processing and can run a large dataset in a fraction of the computation time depending upon computational power available. Another unique advantage of the random forest based QTL method is it's capability to work before a genetic map was created. However, without a genetic map, markers selected by RFQTL will not be able to be positioned within the genome, making comparisons with other QTL studies more difficult.

The use of machine learning as a tool for QTL detection has been proven in this research to be as accurate as conventional QTL analysis methods to identify QTLs. Furthermore, the random forest approach allows for the inclusion of meteorological data to study the effect of age and maturity rate on flowering time in *Miscanthus.* It should be stated however, that there are alternative QTL detection method which could also utilise addition data, such as mixed models. The ability to include addition data such as meteorological information is particularly important for the perennial species such as *Miscanthus* as crops which are likely to be kept in the field for at least ten years without replanting. If a controlled experiment can be created with controlled water availability,

temperature and soil type, this analytical tool can be further exploited to understand the complex relationship between the genotypes and their interaction with environment.

4 Machine Learning for Genotyping–by-Sequencing (GBS) Data Analysis

4.1 Introduction

Genotyping is a process of determining the genetic make-up (genotype) of an individual from DNA sequence. Recent advances in next generation sequencing (NGS) technology have resulted in improvements in both the speed and amount of sequence available for genotyping.

Genotyping-by-sequencing, GBS, is a high-throughput and economical method for creating large numbers of potential genetic markers using next generation sequencing technology (Elshire *et al.*, 2011; Poland & Rife *et al.*, 2012). It is now feasible to use GBS to generate high density markers from species with large complex genomes. GBS is a simple, reproducible, highly multiplexed technique based on the Illumina® sequencing platform (Elshire *et al.*, 2011). However, for most GBS experiments the number of attributes (genetic markers) are much greater than the number of observations and this presents problems for conventional statistical analysis.

Alternatives to GBS include Restriction-site-associated DNA (RAD) tags or Genomic reduction based on restriction-site conservation (GR-RSC). Cronn *et al.* (2012) discussed the three methods and concluded that GBS was the simpler technique because it does not require many of the steps involved in the other approaches, such as size selection. GBS is a cost-effective and efficient way to generate high density SNP markers. GBS correctly identified SNPs related to traits in cultivated barley but the data required more complex analysis than other methods (Liu *et al.*, 2014). The GBS method is suitable for population

studies, germplasm characterization, genetic improvement and trait mapping (Deschamps *et al.*, 2012; Poland *et al.*, 2012; Narum *et al.*, 2013). GBS can be used for Genome Wide Association Study (GWAS) and Genomic Selection (GS) (Meuwissen *et al.*, 2001; Cockram *et al.*, 2010; Brachi *et al.*, 2011). GWAS uses linkage disequilibrium to predict which genomic region(s) influence important traits, while GS predicts desirable phenotypes by calculating breeding values based on genotyping information. The success of both GWAS and GS are highly dependent on the effectiveness of the computational tools used to link markers to traits of interest. Therefore effective methods for data analysis must be developed in order to exploit the data efficiently.

In this chapter, I describe the application of machine learning to the analysis of datasets generated from GBS in *Miscanthus*. A method was developed that utilised machine learning to detect SNP-trait associations. *Miscanthus* requires three years to reach maturity and therefore conventional breeding by phenotypic selection may take a long time. Therefore the potential for marker-assisted selection (MAS) to improve throughput in a *Miscanthus* breeding programmes is significant. Markers must first be identified for traits which are of interest for breeding. The machine learning algorithm random forest was applied to a collection of wild *Miscanthus* germplasm. No physical map exists for *Miscanthus*, therefore markers were mapped to the *Sorghum* genome, to validate potential markers where QTL have been identified in *Sorghum* linked to a given trait. The methodology developed allowed for fast and simple detection of trait associations in GBS data. In the future, we aim to apply this experience for GWAS and GS studies and genetic characterization of populations.

4.2 GBS and Molecular Plant Breeding

4.2.1 GBS and marker-assisted selection

The use of genetic information has long been promoted as a tool for increasing the efficiency of selection in plant breeding (Bernardo, 2001; Bernardo, 2008). Marker-assisted selection (MAS) from QTL mapping has been applied in many crop improvement programmes (Prasanna *et al*., 2010; Steele *et al*. 2013; Ashraf & Foolad, 2013). The identification of markers for MAS is expensive and requires mapping families, phenotyping, sequencing and a lengthy process of marker analysis and QTL detection. The effectiveness of MAS depends on the closeness of the marker-to-trait association and the breadth of genome coverage produced by GBS, in common with other NGS-based techniques, will allow closer marker associations.

Recently, new MAS methods such as GWAS and GS have been used more routinely. For example GS has been used in cattle breeding (Luan *et al.,* 2009; Hayes *et al.,* 2009), ryegrass breeding (Hayes *et al.,* 2013) and in wheat breeding (Sorrells *et al*. 2011); GWAS has been used in humans (Visscher *et al.,* 2012), Rice (Huang *et al.,* 2012), Maize (Kump *et al.,* 2011; Tain *et al.,* 2011) and Barley (Pasam *et al.,* 2012). These studies demonstrated that selection based upon whole genome methods is a potentially powerful tool to improve the efficiency of breeding programmes. Conventional statistics-based marker analysis methods were sufficient when marker numbers were low, such as the markers generated from AFLP studies. However, the information generated from NGS-based techniques such as GBS is massive and complex and new methods for data analysis are needed.

4.2.2 How GBS facilitate Miscanthus breeding

Miscanthus, native to Asia, is an undomesticated, perennial grass. IBERS has one of the largest wild germplasm collections of Miscanthus outside Asia. The breeding programme at IBERS, Aberystwyth University has been breeding high yielding, stress tolerant, seed propagated *Miscanthus* varieties since 2006.

Miscanthus has been reported to display a wide phenotypic variation in its germplasm (Robson *et al.,* 2013; Slavov *et al.,* 2014). This provides breeders with a vast pool of different traits to use in breeding. However there are challenges in breeding *Miscanthus*. *Miscanthus* is self-incompatible, thus, inbred plants with reduced genotypic variation that are used in wheat and maize breeding programmes are not available for genetic studies of *Miscanthus* (Hirayoshi, 1955).

Miscanthus is a perennial crop and is productive for 20 years or more (Clifton-Brown *et al*., 2001; Gauder *et al*., 2012). It is estimated that an establishment period of approximately three years is needed before the crop reaches maximum attainable yield. Although this may vary by environment with warmer climates believed to reach maturity faster. Therefore it will take several years to evaluate yield in M*iscanthus*, whereas in annual crops this would be much faster. All this means a *Miscanthus* breeding programme takes longer and more resources are needed to produce a new variety. The urgent question is ‘how to reduce the evaluation cycle?’.

To meet these challenges, the application of new technologies, such as GBS, are vital to accelerate the domestication process of *Miscanthus* in a cost-effective manner. It is therefore desirable to create models based on genomic information generated by GBS to provide a prediction of potential yield from genotype; such models could be used to

evaluate new progeny thereby reducing the time taken to evaluate new genotypes.

4.3 The Power of Machine Learning Approach on GBS Data Analysis

Datasets generated from GBS as mentioned before are high dimensional in that the number of attributes is much greater than the number of observations. Machine learning has already been shown to be effective in genomic selection studies which also utilise high dimensional data sets (Heslot *et al*., 2012) with methods such as support vector machines, neural networks and random forest being applied.

The analysis of GBS data is a complex process. It is highly unlikely that the whole genome of an organism would be responsible for any given trait, in fact it's likely to be many small effects across the whole genome with potentially a few large effects in a smaller number of regions (Buckler *et al*., 2009). Therefore any method used must be able to pinpoint markers in order to find those related to the trait.

Noise in data could also present a problem in GBS studies, missed calls for alleles, errors in phenotyping either caused by measurement error or by low frequencies of measurements can result in these complex interactions being missed.

To efficiently and correctly model associations in GBS-like studies methods are needed that are noise tolerant, able to perform attribute selection and are capable of handling high dimensional data.

Machine learning is well-known to possess all three of these qualities. Random forest (Breiman, 2001a), a machine learning algorithm that involves the creation of many decision trees, uses bagging and bootstrapping to deal with noisy data. In Chapter 3, it was demonstrated that the importance scores created in a random forest model allow for the identification of markers which are affecting the trait being modelled in a mapping

Figure 4.1: Distribution of the Miscanthus collections used in this genotyping-by-sequencing analysis showing a diverse distribution of latitude and longitude. Image generated using http://www.gpsvisualizer.com.

family.

In the following section a study of GBS datasets will be presented from studies of *Miscanthus* using random forest to search for markers that relate to traits of interest.

4.4 Results and Discussions

4.4.1 Genotype Selection

From a diverse germplasm collection available at IBERSm 244 *Miscanthus* genotypes were selected for GBS analysis from a wide geographical distribution with latitude range between 18º to 45º N. The selected genotypes also came from a range of altitudes between -11m and 2.5km above sea level. The distribution was plotted onto Google Earth as shown in Figure 4.1. The chosen genotypes include *M. sinensis*, *M. sacchariflorus*, hybrids and *M. floridulus* along with some other *Miscanthus* species. Plants were grouped

Breeding classification	Count	Percentage Dead
M. condensatus	14	14.28
M. floridulus	28	39.28
M. lutarioriparius	58	10.34
M. robustus	12	16.66
Sacc/sin	16	6.25
M. sacchariflorus	148	10.13
M. sinensis	183	16.39
Hybrid	16	0

Table 4.1: Survival rates of several Miscanthus species included in the GBS trial

together into their breeding classifications for analysis, these are the suspected 'species' classification given to a genotype at time of collection based upon morphological analysis.

Out of the 244 genotypes that were selected for GBS, at the time of writing only 179 had available complete marker data. 3778 bi-allelic SNP markers were generated. Each was recoded as either A or B for the homozygotes H for the heterozygotes, and 0 indicated where no call was detected for that genotype. Each marker was aligned against the *Sorghu*m genome and the position of the SNPs were recorded.

4.4.2 Trial planting and survival

Two replicates of each genotype were planted in northern Germany at the Julius Kühn-Institut (JKI) in April 2013 by colleagues from the institute. The trial was irrigated after planting to aid establishment. I then collected phenotype data in October 2013, with 6 observations, canopy height, tallest stem, stem count, base diameter, stem diameter, and flowering score, were collected on all surviving plants using the protocols outlined in chapter 2. The yield data of the first year’s harvest was taken in March 2014.

Out of the total 475 plants, 67 did not establish. The species distribution of the

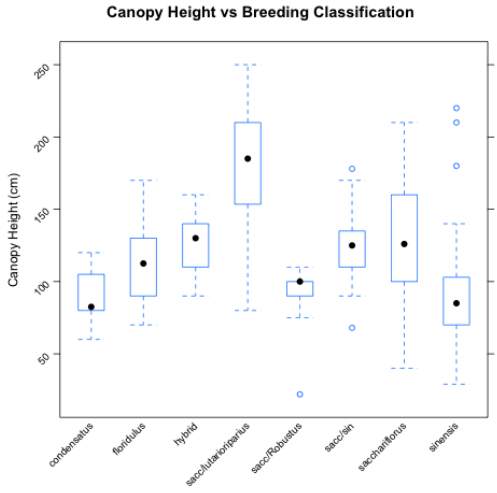

Figure 4.2: Canopy height from different Miscanthus breeding classifications grown for 1 year at a field site in Northern Germany October 2013.

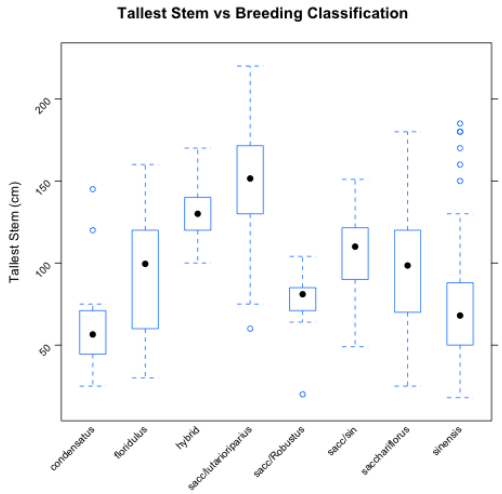

Figure 4.3: Tallest stem from different Miscanthus breeding classifications grown for 1 year at a field site in Northern Germany October 2013.

surviving and dead plants is shown in Error: Reference source not found. Species classifications used in the table come from the group database MSCAN and are assigned at collection by the scientist collecting the accession. As more data comes available these maybe corrected. Any genotypes that are made by the group and are between two different species are known as hybrids, whereas those which are suspected to have occurred in the wild are know as Sacc/Sin. *M. floridulus* had the highest fatality rate with almost 40% of the plants failing to establish.

4.4.3 Phenotyping

A wide range of different canopy heights were observed from the highest at approximately 2.5 metres (*M. lutarioriparius species)* to as little as 20 centimetres (Figure 4.2). Observations of the tallest stem were taken and again *M. lutarioriparius* had on average the tallest stem (Figure 4.3) The tallest stem usually extends beyond canopy height in flowering plants (Figure 4.4). Not all of the genotypes flowered, this could be due

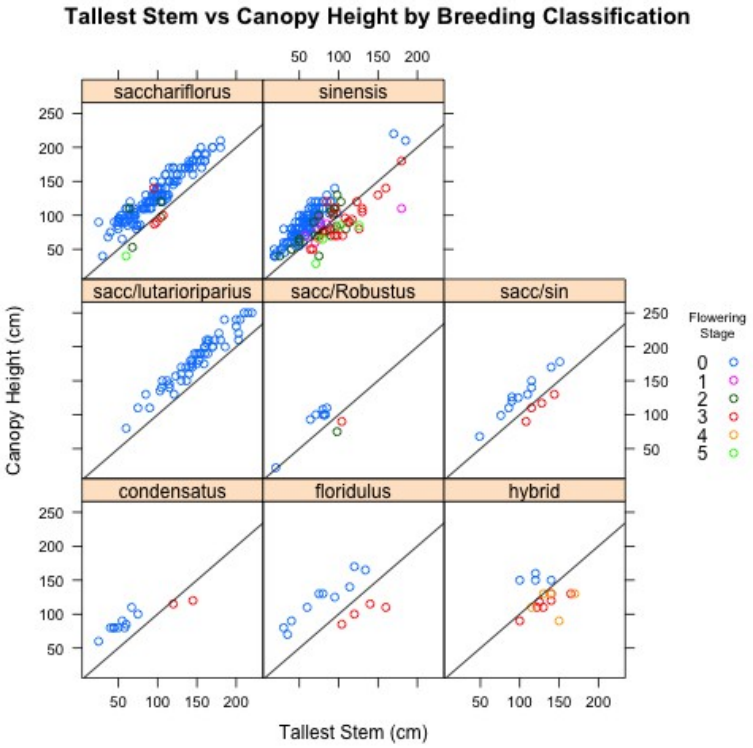

Figure 4.4: Canopy height plotted against the tallest stem measurements. The line shows when the two measurements are equal. Non flowering genotypes (flowering stage 0) lie above the line, whereas the flowering genotypes (flowering stage 1 – 5) tend to lie below the line.

to immaturity of the plant. One theory to explain the lack of flowering is that it is the first season's growth and development is focussed on increasing biomass to improve the chance of surviving the first winter. Therefore a plant may not flower to save expending energy and resources. However, it may be due to the fact that the environment was not in a suitable condition to induce flowering in some genotypes even when mature, further observations are needed to confirm this.

The hybrids had the highest stem counts (Figure 4.5) but these stems were on average thinner than stems observed in any other species groups (Figure 4.6). In general, *M. lutarioriparius* had the thickest stems with an average of approximately 10 mm and a maximum of 15 or 20 mm.

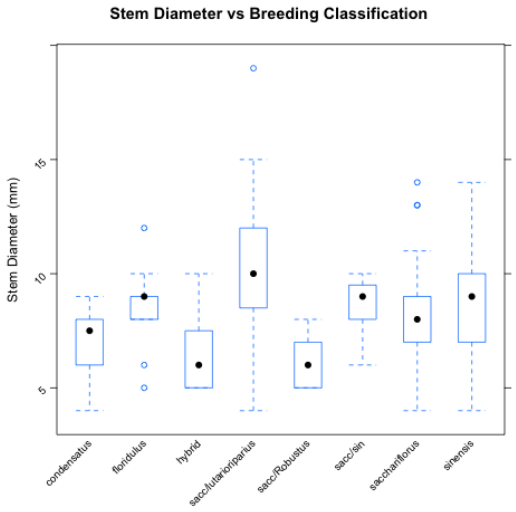

Figure 4.5: Stem count from different Miscanthus breeding classifications grown for 1 year at a field site in Northern Germany October 2013.

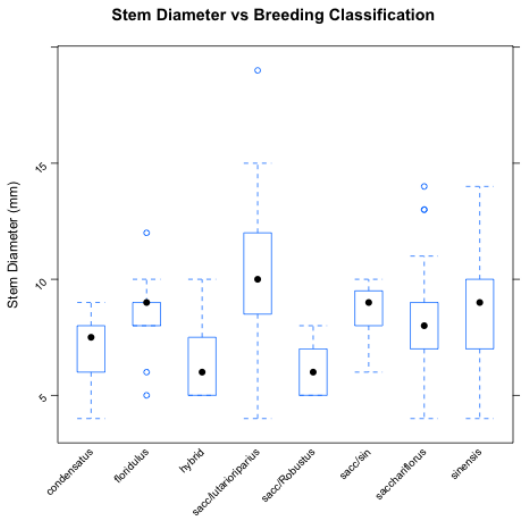

Figure 4.6: Stem diameter from different Miscanthus breeding classifications grown for 1 year at a field site in Northern Germany October 2013.

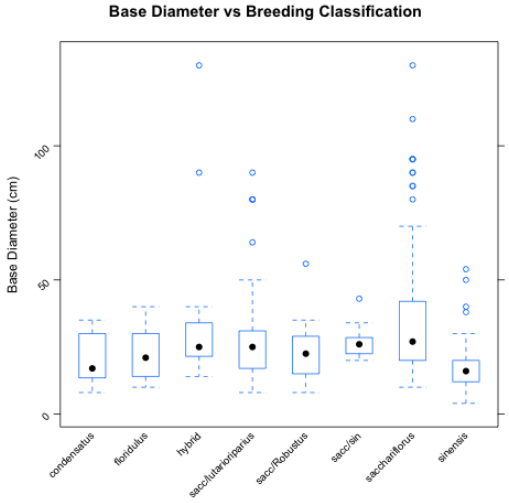

Figure 4.7: Base diameter from different Miscanthus breeding classifications grown for 1 year at a field site in Northern Germany October 2013.

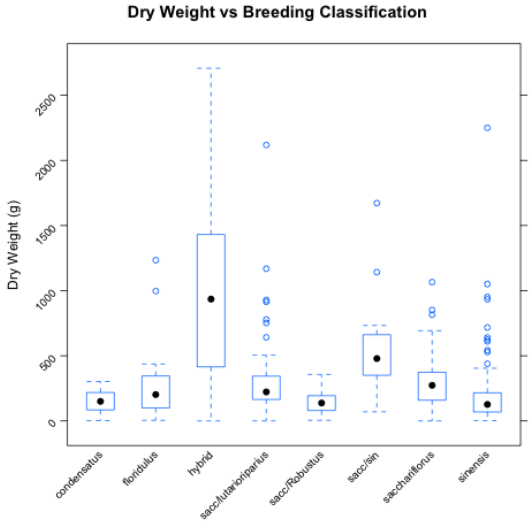

Figure 4.8: Dry weight from different Miscanthus breeding classifications grown for 1 year at a field site in Northern Germany March 2014.

As expected *M. sacchariflorus* had the highest average base diameters given their creeping nature (Figure 4.7). Some hybrids also displayed creeping characteristics. Creeping could be an important trait potentially leading to gap filling in commercial trials between plants. This would be most effective if plants did not leave gaps when creeping, unlike *M. x giganteus* which leaves a hollowed out centre as it grows and creeps.

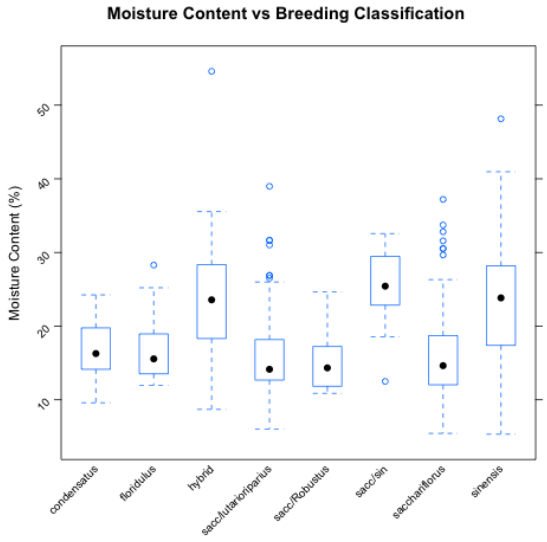

Figure 4.9: Moisture content from different Miscanthus breeding classifications grown for 1 year at a field site in Northern Germany March 2014.

The plants were harvested in March 2014. The dry weight (Figure 4.8) and moisture content (Figure 4.9) of each plant were measured. The hybrids displayed the highest yield. The highest moisture content was found in the hybrids, the sacc/sin group and the sinensis group. The sinensis group also displayed a large variation in the observed moisture contents.

4.4.4 Phenotype Comparisons

To compare the relatedness of the various phenotypes, observed correlations plots were generated for all plants (Figure 4.10). The numbers displayed in the boxes of Figure 4.10 are the R squared values from fitting a linear model between the pairs of variables. The best predictor of yield (dry weight) is stem count. This differs from other studies where canopy height was the best predictor of yield (Robson *et al*., 2012). Differences in trial design could be the cause of this discrepancy. In Robson *et al*'s study they did not use stem counts in their model, instead they used transect count. This is a measurement they

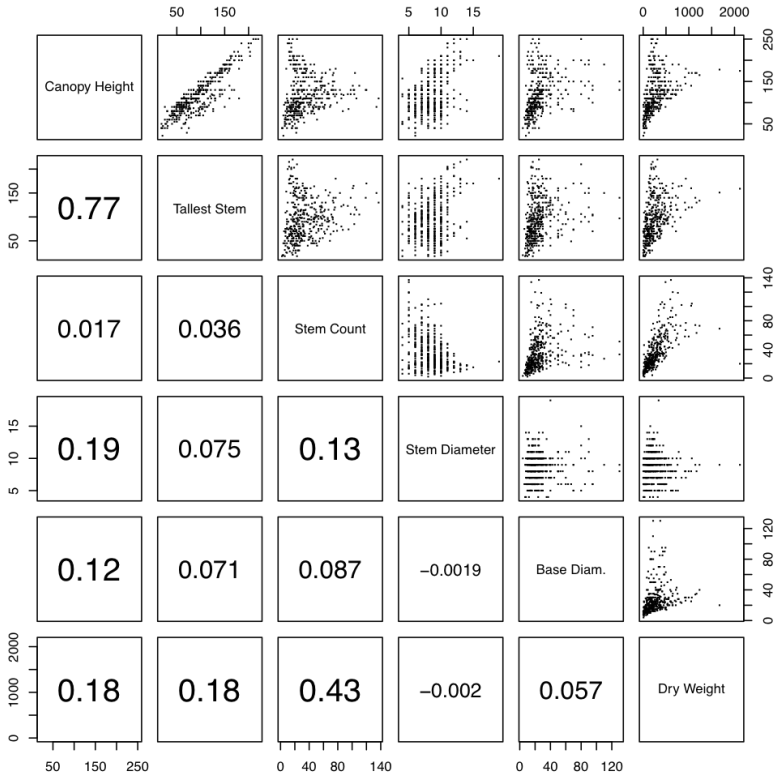

Figure 4.10: Grid shows the R squared values from a linear model built using the R lm() function. Stem count has the highest correlation to the dry weight.

used to estimate the stem count by passing a stick through the middle of the plant and seeing how many stems touched the stick. However if this measure did not accurately reflect the stem count this could explain the differences between the two models.

Apart from the high correlation between yield and tallest stem or canopy height, few other traits had significant correlations with yield. However every species displayed different combinations of morphologies, thus correlations may be studied within species.

Pairwise linear models were also used to calculate the R-squared values between traits within each breeding classification (Figure 4.11). For the *M. sinensis* classification stem count had the strongest association with stem count. Yield in *M. sinensis* had strong

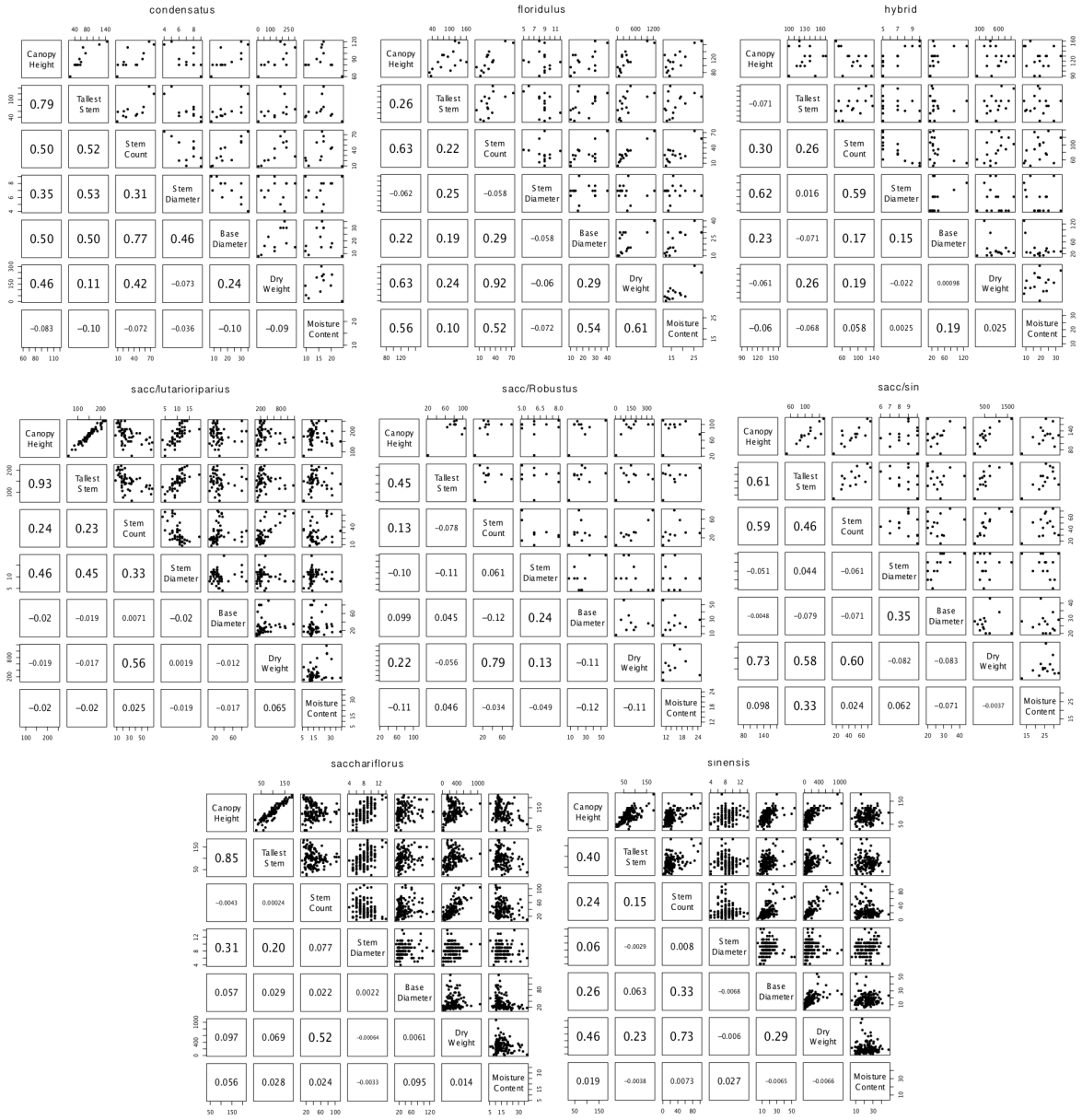

Figure 4.11: Each grid shows the R squared values between several phenotypic traits of 8 Miscanthus species.

association with stem count. Canopy height was also shown to associate with yield however this was lower than stem count for *M. sinensis.* Considering *M. sacchariflorus* types, including its closely related species *M. lutarioriparius* and *M. robustus,* stem counts are the best predictor of yield. There was little correlation between any phenotypes and yield in the hybrids. Only a small number of hybrids were in this trial, so there is less chance that strong correlation can be found. The genotypes classified as Sacc/Sin showed

strong association between yield with canopy height and stem count. Tall stems will make a better yielding plant. This is reflected in the morphology of *M. x giganteus s*een in the naturally occurring cross between M. *sacchariflorus a*nd M. *sinensis.*

4.4.5 Species Classifications

Genetic dissimilarity was calculated in order to understand the genetic relationship between species. Dissimilarity was calculated as the average differences of the sum of marker scores described below. Each marker that exists in both genotypes was compared. If the allele was the same in both markers, 0 was added to the running total. If either one contained a heterozygote and the other was a homozygote, 0.5 was added to the running total. Finally if the two makers were different homozygotes, then 1 was added to the running total. The total was then divided by the number of marker pairs compared to give the genetic dissimilarity of the two genotypes. A matrix of dissimilarities between all genotypes in this GBS analysis was created as illustrated in Figure 4.12. Multi-dimensional scaling was used to visualise the dissimilarity matrix using the cmdscale function in R.

Genetic dissimilarity scores divided *Miscanthus* into two main groups, *M. sinensis* with its adjacent species *M. floridulus* and *M. sacchariflorus* with its related species such as *M. lutarioriparius*. In the middle between the two groups is *sacc/sin*. Given they are the assumed hybrids of two main species, they would be expected to lie between the two. *M. x giganteus* also appears in central area which is consistent with the finding of Hodkinson *et al* (2002b) that *M. x giganteus* is a hybrid between *M. sinensis* and *M. sacchariflorus*.

From the results of this analysis, several genotypes were identified as possibly being incorrectly classified. Several genotypes that had been classified previously as *M. floridulus* are shown to be more genetically similar to *M. sacchariflorus.* Chou (2009)

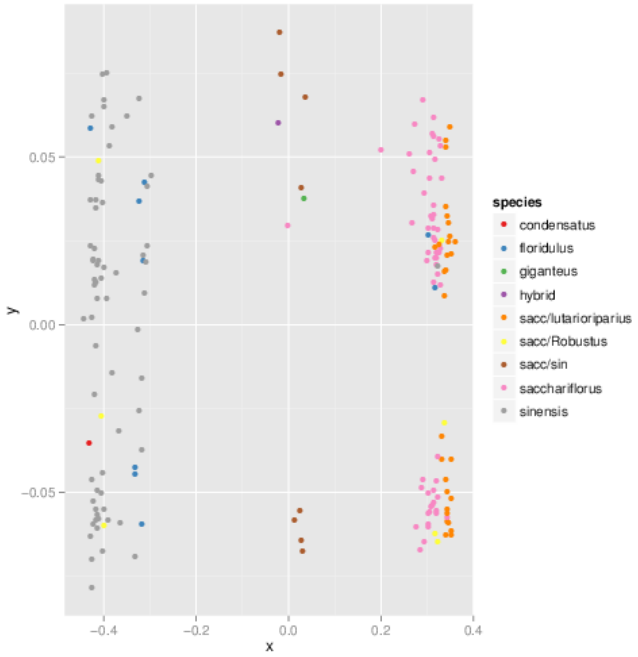

Figure 4.12: Multidimensional scaling of the genetic dissimilarity between all genotypes in the GBS study. Miscanthus appears to have three groupings, the left hand side is mainly M. sinensis and M. floridulus, the right hand side are M. sacchariflorus like genotypes (including M. lutarioriparius) and in between lie the sacc/sin plants.

showed that *M. floridulus* has evolved from *M. sinensis*. Therefore it is less likely that *M. floridulus* would be genetically more similar to *M. sacchariflorus* than *M. sinensis*. This suggested that they may have been miss-classified upon collection. After morphological re-evaluation within the breeding programme the decision was made to reclassify these genotypes as *M. sacchariflorus*.

4.4.6 Trait Associations

In order to investigate markers that associate with traits of interest for breeding, the phenotype data and genetic markers were analysed using machine learning. Random forest (Breiman, 2001a) was used to look for markers that show high importance for traits. Random forest was used to perform regression analysis of the markers with the observed trait. A default mtry parameter value of p/3, where p is the number of markers, was used

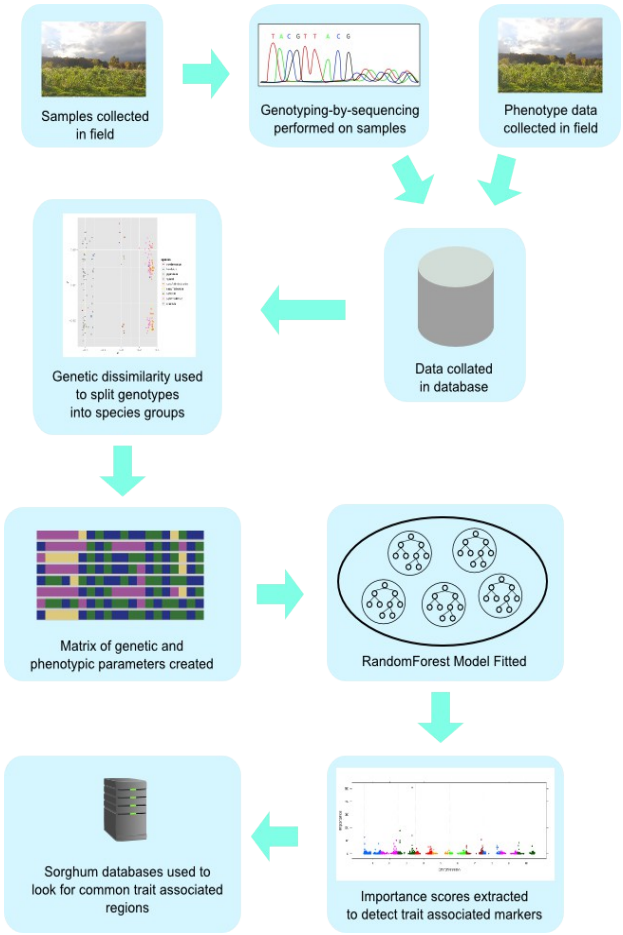

Figure 4.13: Flow diagram describing how analysis was performed described in this chapter. First samples were collected in the field and GBS was performed. Phenotypes were also collected from a trial in JKI. This data was integrated into a local database. From this genotypes were classified using genetic dissimilarity into two groups. These groups where then converted into a matrix of parameters and observations. They were analysed using random forest algorithm. Importance scores were extracted and comparisons were made with Sorghum to look for traits associations.

and 500 trees were trained in each forest. Data was analysed using the pipeline described in Figure 4.13.

Importance scores were used to detect SNPs which related to the traits of interest. Importance scores are calculated as average difference of the out-of-bag error before and after permutation over all the trees. The importance score ranks markers, those with the highest importance are having the greatest effect on the response variable, in this case the trait being modelled. Although any marker with high importance has a strong effect on the

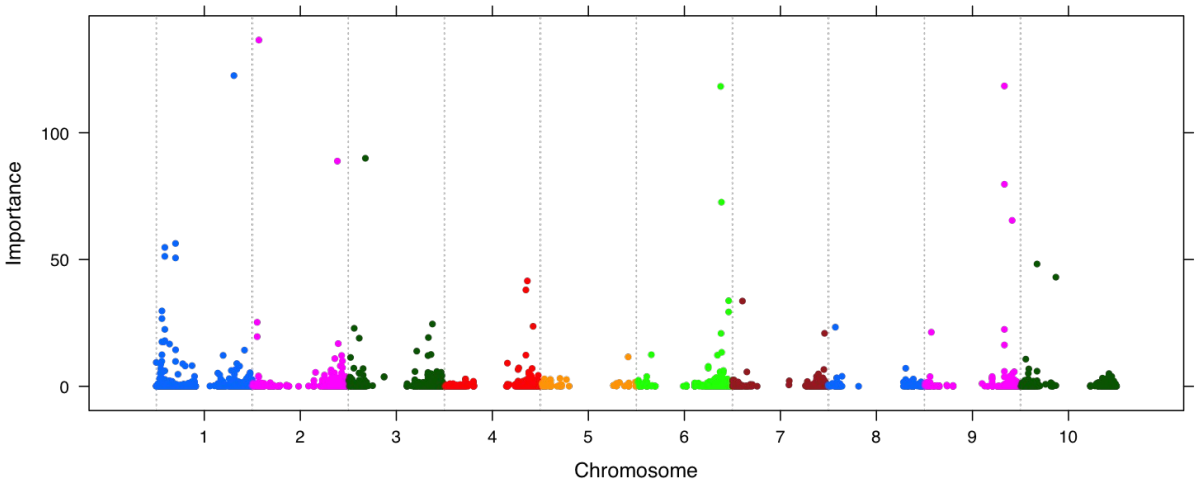

Figure 4.14: Importance scores from a Random Forest model of the GBS data for canopy heights in Miscanthus. As there is no published physical map of the Miscanthus genome the markers were mapped to the Sorghum genome. The model explained 74.67% of the variance. Important markers were seen across several chromosomes including 1, 2,3, 6 and 9.

trait. The most important markers will be investigated to look for evidence of this marker related to the trait of interest to test random forests ability to detect related markers. Where relevant SNP's were identified the CSGR-QTL database (Zhang *et al*., 2013) was queried for possible *Sorghum* QTL that relate to the markers found by the Random Forest association method. Markers were mapped to the Sorghum genome (Paterson *et al.,* 2009) as there is no published *Miscanthus* physical map. 13 markers could not be mapped to *Sorghum*, however none of these had high association with any of the traits investigated.

Canopy Height

A random forest model was fitted for canopy height prediction. The model explained 74.67% of the variance. High importance markers were seen on chromosomes 1, 2, 3, 6 and 9 (Figure 4.14). The most important marker aligned to Chromosome 2 of *Sorghum*. The markers on chromosomes 1, 6, 9 displayed similar levels of importance. The phenotypic distribution for the most important marker showed a vast variation in heights

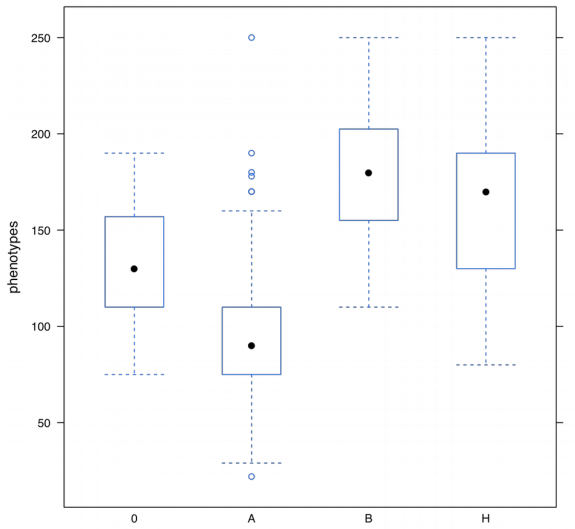

Figure 4.15: Highest importance marker for canopy height mapping in the Miscanthus genotypes. Alleles B and H appear to be in association with the tallest genotypes.

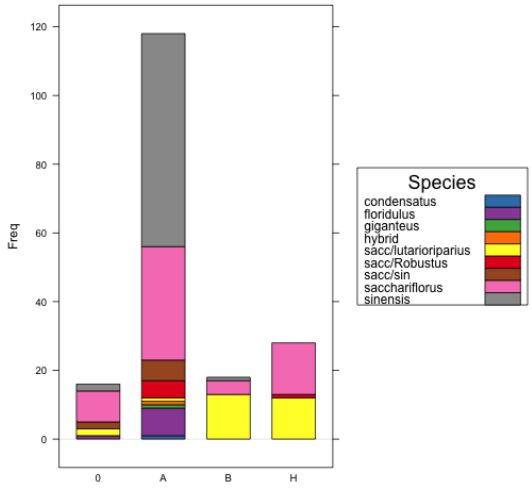

Figure 4.16: The species distribution for the highest importance marker for canopy height. The B and H alleles seem to only exist in M. lutarioriparius and M. sacchariflorus.

between the two homozygotes (Figure 4.15). The B and H alleles were shown to associate with taller canopy heights. When we examined the species classifications the B and H alleles were almost exclusively found in *M. lutarioriparius and M. sacchariflorus* (Figure 4.16).

Based on the data illustrated in Figure 4.2 it is obvious that the *M. lutarioriparius* genotypes had a higher canopy height. Therefore it would be appear that random forest is selecting markers that can distinguish species. The same distribution is also seen in the other high importance markers. Random forest analysis appeared to identify markers that split observations into the species groups. Although once species have been split it would appear to detect markers that explain the differences in height. This therefore makes it more difficult to know which markers are related to canopy height.

Species Groupings

Given this problem it was then decided to split the data into two subsets. This splitting

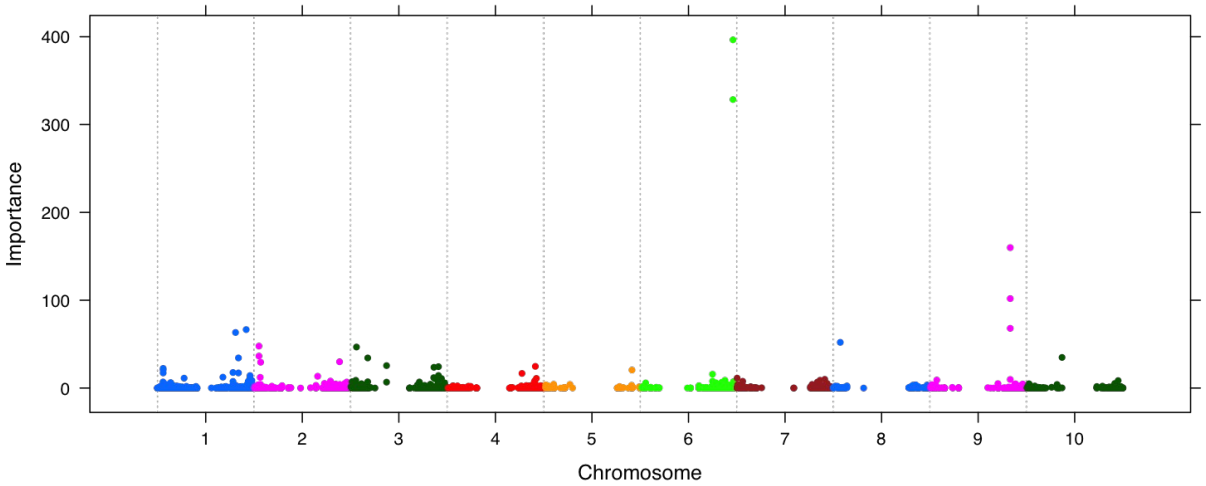

Figure 4.17: Importance scores from random forest modelling of canopy height for a M. sacchariflorus subset. The model explained 66.97% of the variance in the canopy height observed. Present are high importance peaks on chromosome 6 and 9. Other markers appear significant but not as high as the two previously mentioned.

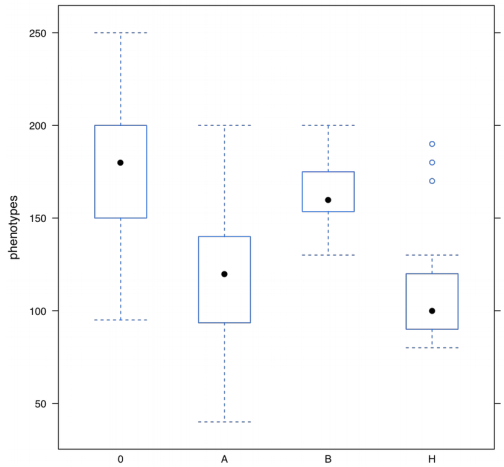

Figure 4.18: The most important marker of canopy height for a M. sacchariflorus subset. Without this marker plants are generally taller, although the B allele Is the more favourable one for height

was selected using the data form the multidimensional scaling done using genetic dissimilarity (Figure 4.12). The two groups were selected from either side of the graph. The left group will be referred to as the *M. sinensis* subset. The right group will be referred to as the *M. sacchariflorus* subset. The genotypes which appear between these two groups, that have the species classification of either *M. x giganteus* or *sacc/sin* were put into a third group which was not analysed due to low numbers.

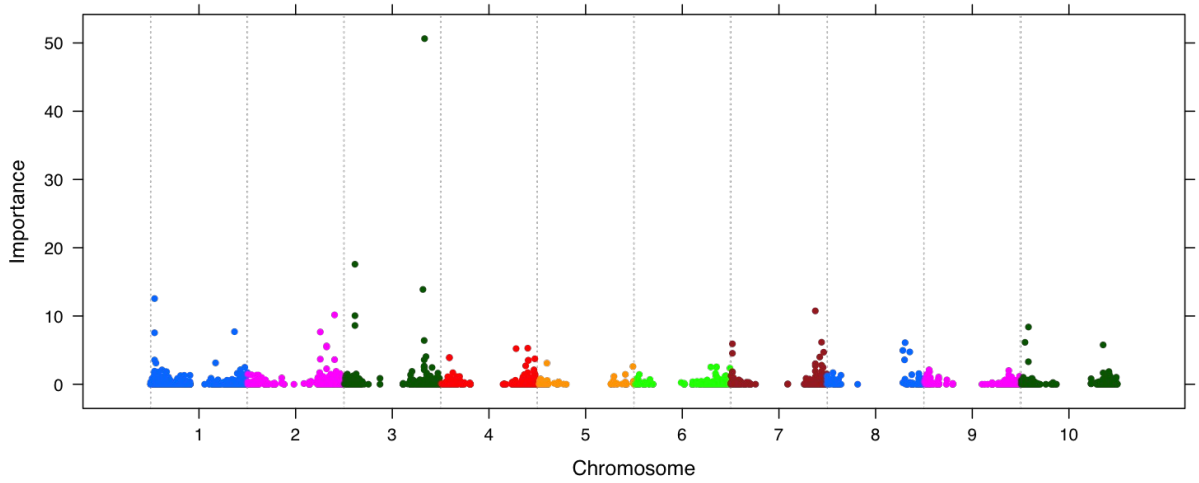

Figure 4.19: Importance scores from random forest modelling of canopy height for a M. sinensis subset. The model explains 43.03% of the variance seen in the M. sinensis canopy heights. A high importance marker is seen on chromosome 3 when mapped to Sorghum. Potential other important markers are seen on chromosome 1, 2, 7, 8 and 10.

M. sacchariflorus subset

The random forest model was fitted to the *M. sacchariflorus* subset. The model explained 66.97% of the variance. We saw two high importance regions, one which was mapped to *Sorghum* Chromosome 6 and the other to Chromosome 9 (Figure 4.17). The two SNP markers on Chromosome 6 are co-located in the *Sorghum* genome. The lack of this SNP associates with a taller canopy height. One possible explanation is that the SNP is located in or linked to a dwarfing gene (Figure 4.18). The marker lies within a documented plant height QTL of *Sorghum* (Shiringani *et al.,* 2010) (Chr6: 58257387-62208784). No match for an associated plant height QTL was found for the SNP on chromosome 9.

M. sinensis subset

The model fitted for the *M. sinensis* subset explains 43.03% of the variance observed in canopy height. A high importance marker is seen on chromosome 3 when mapped to *Sorghum* (Figure 4.19). Other potentially important markers are seen on chromosome 1, 2,

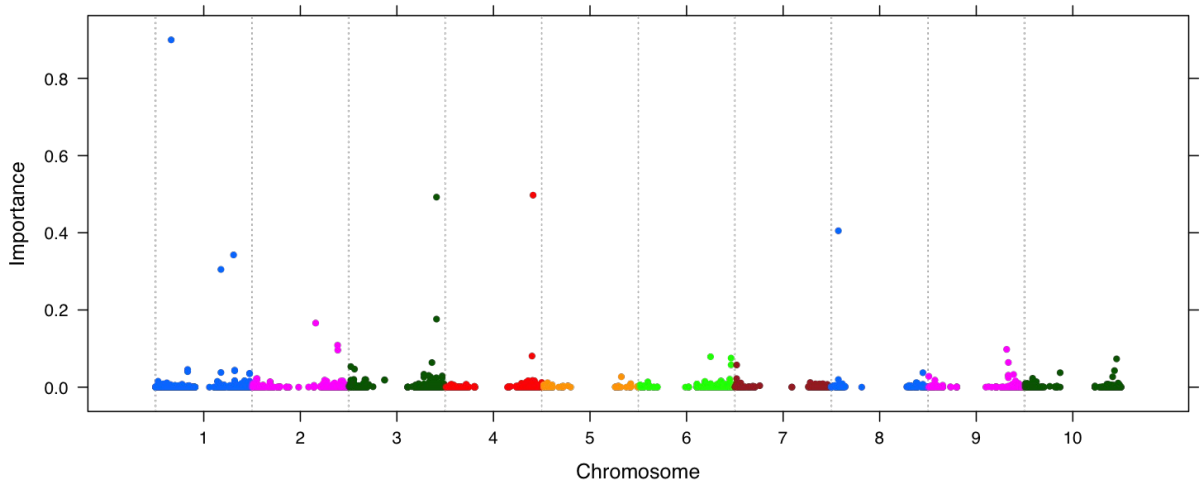

Figure 4.20: The importance scores from the random forest model of stem diameter for a subset of M. sacchariflorus subset. The model explains 56.51% of the variance seen in the stem diameters observations. Important markers are seen on Chromosome 1, 3, 4 and 8 after the markers have been mapped onto the Sorghum genome.

7, 8 and 10. The trait related marker as detected by random forest analysis seen towards the end of chromosome 3 lies within three documented plant height QTL PTHT-3-2, PTHT-3-1 (Ritter *et al.,* 2008) (Chr3: 55215143 – 68161815; Chr3: 53558698 – 68161815) and HtAvg-3-1 (Lin *et al.,* 1995) (Chr3: 55866462 – 67437541). The trait related marker detected on chromosome 7 lies within another plant height QTL, PTHT.7b (Chr7: 54201209.0 – 61172136.0). The related marker on chromosome 1 and at the beginning of chromosome 3 do not match to any documented plant height QTL.

Stem Diameter

M. sacchariflorus subset

Within the *M. sacchariflorus* subset the random forest model explained 56.51% of the variance associated with stem diameter. Important markers were detected on Chromosome 1, 3, 4 and 8 (Figure 4.20). Little study has been performed on stem diameter in *Sorghum*. No QTL were found in the CSGR-QTL database. Although one study has suggested that QTL exists on chromosomes 4 and 6 (Zou *et al*., 2012).

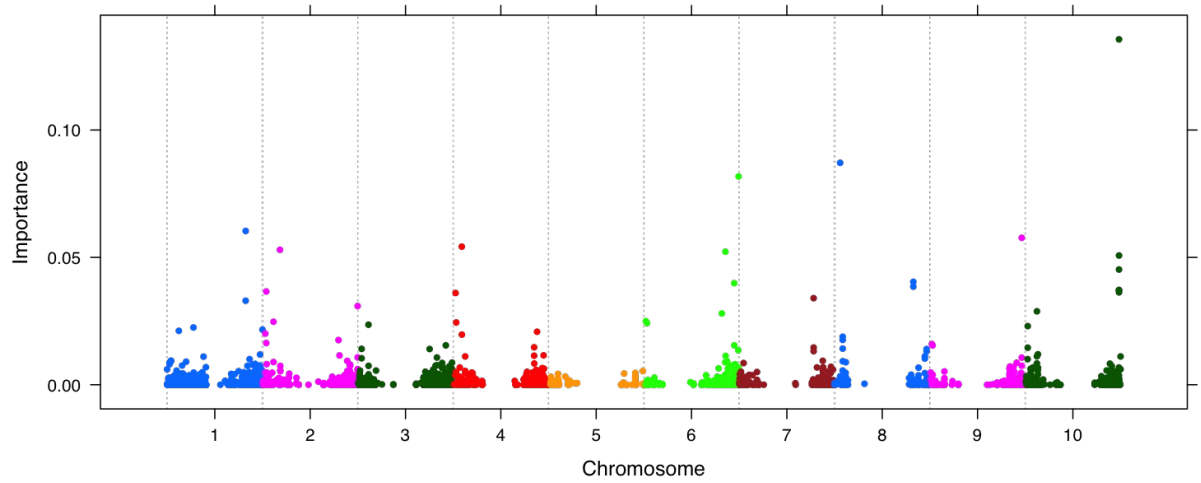

Figure 4.21: The importance scores from the random forest model for M. sinensis subset for stem diameter have been plotted against the Sorghum genome. The model explains 22.5% of the variance. High importance markers were found on chromosome 6, 8 and 10.

M. sinensis subset

The random forest model fitted for stem diameter to the *M. sinensis* subset explained 22.5% of the variance, lower than was seen in the *M. sacchariflorus* subset. Several trait related markers were detected on chromosomes 10, 8, 6 (Figure 4.21). Lesser associations were also found on chromosomes 1, 2, 4. As discussed earlier there are no documented stem thickness QTL in the database, however a study has suggested a potential QTL on chromosome 4 and 6 in *Sorghum* (Zou *et al.,* 2012).

Base Diameter

Base diameter was measured on this population but the variance cannot be explained in either of the datasets. This trait is likely to be driven largely by the maturity of the plant as the *M. sacchariflorus* displays a creeping phenotype or it may be that this trait is better observed in a more mature plant.

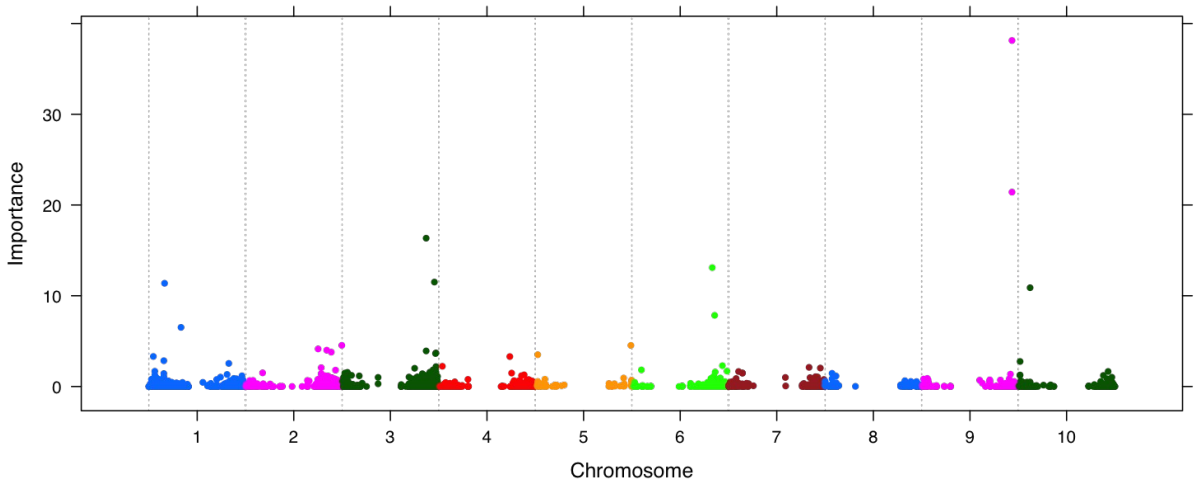

Figure 4.22: The importance scores for stem count in the M. sacchariflorus subset. The model explained 49.99% of the variance seen in the stem count observations. The markers have been mapped onto the Sorghum genome. High importance markers are seen on chromosome 9, 3, 6 and 1.

Stem Count

M. sacchariflorus subset

Stem count was previously shown to be one of the most important traits for predicting yield (Section 4.4.4). The variance explained by the random forest model in the *M. sacchariflorus* subset was 49.99%. Marker related traits were detected by the random forest model on chromosome 9, 3, 6 and 1 (Figure 4.22). The two most important markers are only 10bp apart on chromosome 9, however no stem count QTL has been reported at these loci in Sorghum. One marker related to stem count was detected on chromosome 1 which lies within the documented tiller number QTL Tinb-1-2 (Hart *et al.,* 2001) (Chr1: 52708000 – 58925629). The marker detected on chromosome 6 lies within another tiller number QTL TINB.6-1 (Shiringani *et al.,* 2010) (Chr6: 58257387 – 62208784).

M. sinensis subset

Only 26.06% of the variance associated with stem counts was explained in the *M. sinensis* subset. Markers related to stem count were detected on chromosome 1, 4 and 10

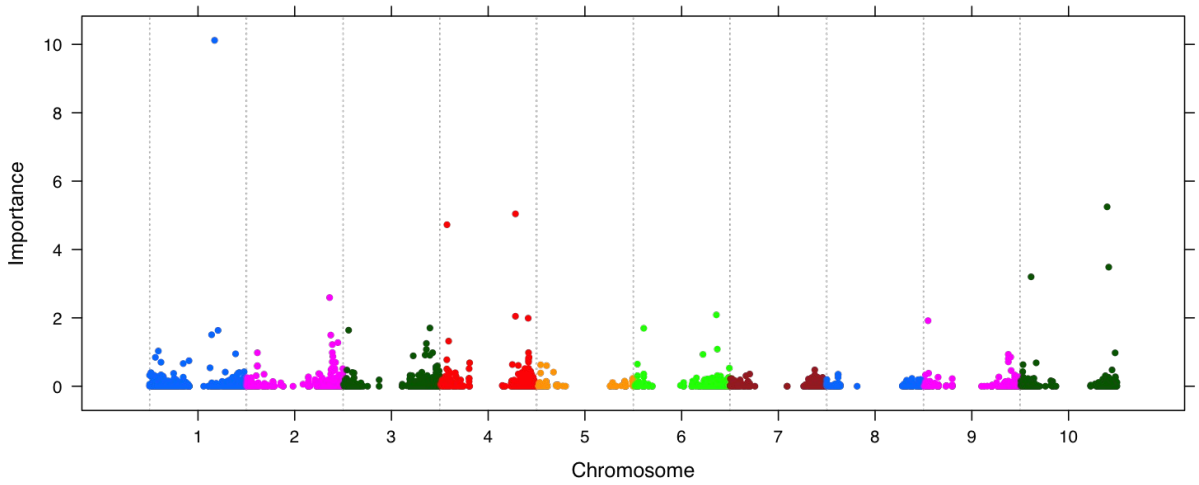

Figure 4.23: The importance scores for stem count in the M. sinensis subset. Only 26.06% of the variance observed in the stem count could be explained by the random forest model. The markers have been mapped onto the Sorghum genome. High importance markers are seen on chromosome 1, 10 and 4.

(Figure 4.23). The most important one was found on chromosome 1. The trait related marker detected on chromosome 1 lies within the tiller number QTL Tillers-1-1 (Paterson *et al.,* 1995) (Chr1: 14162990 – 53606220). The other markers detected on chromosome 10 and 4 do not match to any QTL in the CSGR-QTL database.

Tallest Stem

M. sacchariflorus subset

In the *M. sacchariflorus* subset 64% of the variance was explained by the marker analysis of the tallest stem data. This is very similar to the canopy height models and the same regions of the genome were identified in both analyses. This is not unexpected as the two traits are highly correlated as seen in Figure 4.4 and Figure 4.11.

M. sinensis subset

The variance explained for tallest stem in the *M. sinensis* subset was 33.06%. Related markers were detected on chromosome 1, 3, 4 and 9 (Figure 4.24). When compared to the similar trait of canopy height, additional important markers appeared on

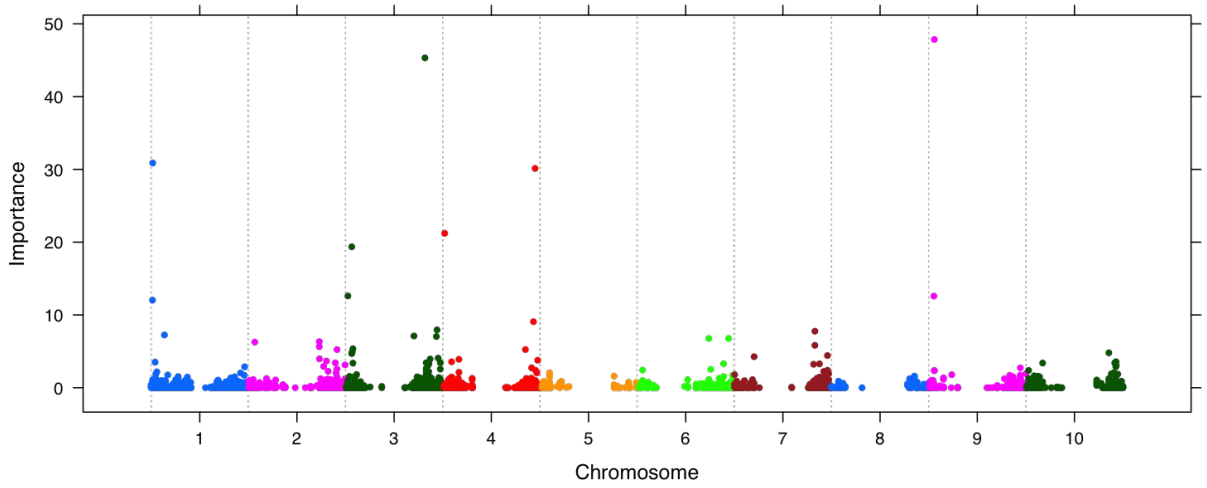

Figure 4.24: Importance scores from the Random Forest model for tallest stem in the M. sinensis subset. It is seen that several of the regions that showed importance for canopy height also appear in this model, but several others are also present. Given that tallest stem is related to canopy height but the relationship differs depending on the flowering stage of the plant it could be these new markers are related to flowering time.

chromosome 1, 4 and 9 (Figure 4.19). Given the relationship between tallest stem and flowering, there is the possibility that this marker could be associated with flowering as it was not identified as important in the canopy height model. However the association on chromosome 9 does not match to any flowering QTL but does lie within a plant height QTL, HtM-9-1 (Lin *et al.,*1995) (Chr9: 571950 – 5164252).

Yield & Moisture Content

The results from the random forest model showed very little variance explained by the markers when modelling moisture content, therefore only yield will be considered.

M. sacchariflorus subset

The yield models explained a variance of 34.11% in the *M. sacchariflorus* subset. The most important markers relating to yield were detected on chromosome 1, 9 and 10 (Figure 4.25). However many other markers were observed across the genome. This suggests that yield is a complex trait which is made up of many genetic interactions. Only four QTLs exist in the CSGR-QTL database responsible for biomass yield. The markes

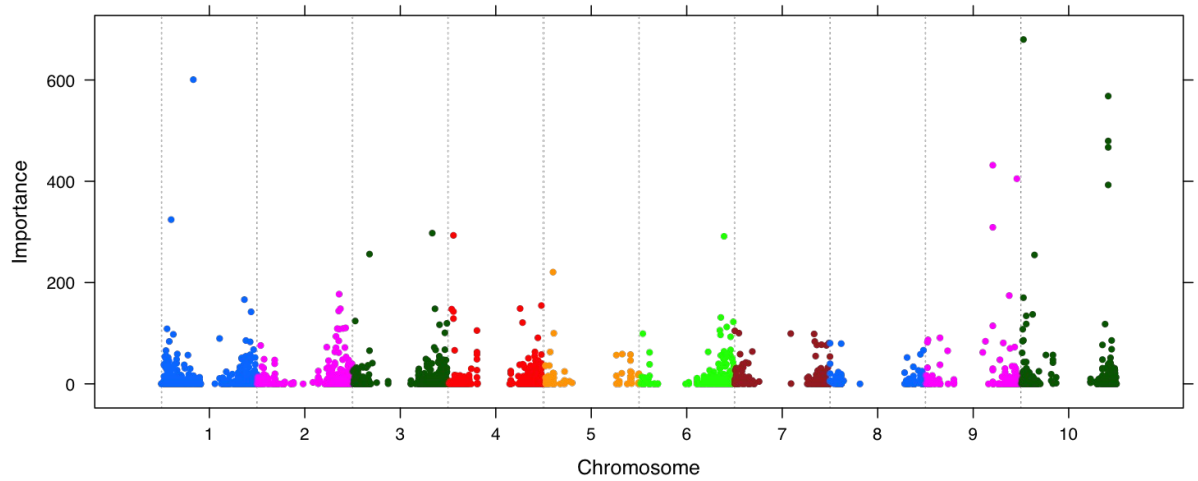

Figure 4.25: The importance scores from the random forest models for yield in the M. sacchariflorus subset. The model explained 34.11% of the variance seen in the dry matter measurements. Important markers were detected on chromosome 1, 9 and 10. Many other markers across the genome appear to correlate to yield, suggesting it is a complex trait which is a composite of many genetic interactions.

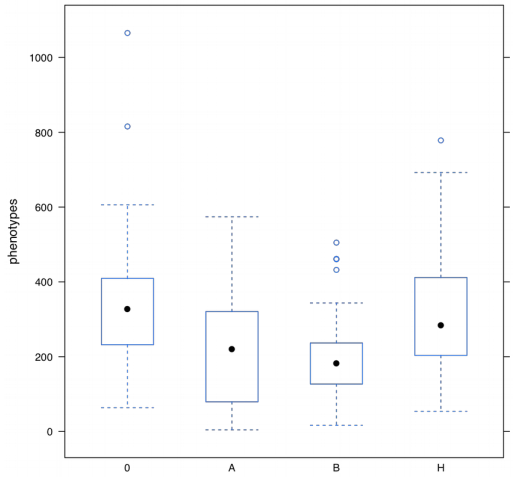

Figure 4.26: The most important marker from chromosome 10 for yield in the M. sacchariflorus subset genotypes selected by random forest. This marker appears to display over dominance. The heterozygote and where the marker is not found it also appears to have a higher yield.

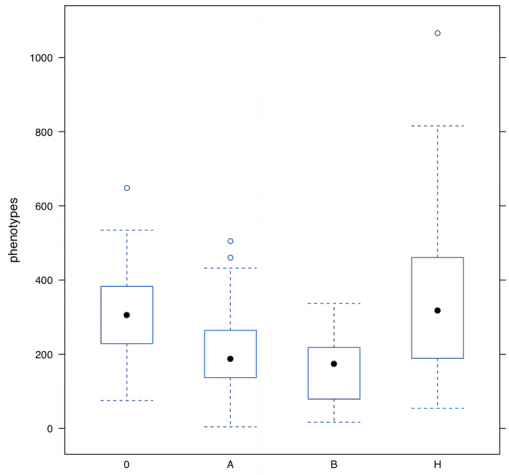

Figure 4.27: The highest importance markers for yield that were mapped to chromosome 1 of Sorghum. We again saw over dominance of the heterozygote.

detected at the end of chromosome 10 lie within one of the yield QTL (Ritter *et al.,*2008) (Chr10: 10996433 – 58245284). Several of the top markers responsible for yield, one located on chromosome 10 and the other on chromosome 1, appear to display overdominance (Figure 4.26 and Figure 4.27). Overdominance is where the phenotype of heterozygote is greater than both the homozygotes phenotypes.

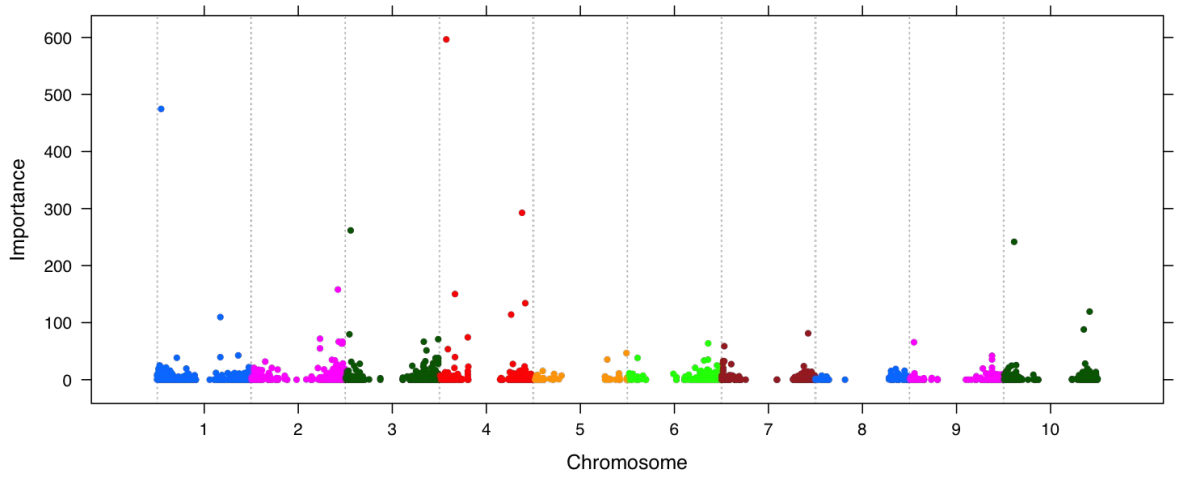

Figure 4.28: The importance scores from the Random Forest models for yield in the M. sinensis subset. The model explains 13.62% of the variance observed. High importance markers were detected on chromosome 4, 1, 3 and 10.

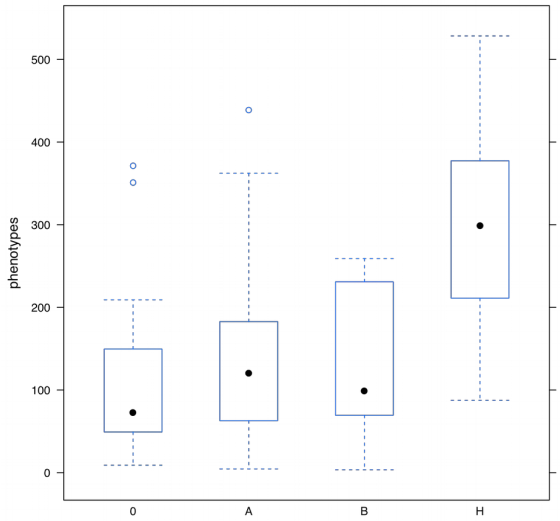

Figure 4.29: The most important markers for yield in the M. sinensis subset as selected by Random Forest .Again the overdominance effect seems to exists in this marker.

M. sinensis like subset

The variance explained by the random forest model in the *M. sinensis* subset was 13.62%. Several markers related to yield were detected on chromosomes 1, 3, 4 and 10 (Figure 4.28). The most important marker for yield was detected on chromosome 4. As with the markers from the *M. sacchariflorus* subset, this marker also appeared to display

over dominance (Figure 4.29).

Discussion

Phenotypic Relationships

To conclude, based on this study, canopy height seems to have a strong association to yield in several of the classifications; sacc/sin, sinensis, floridulus, and condensatus (Figure 4.11).

However, stem count had a stronger association with yield than canopy height within several classifications; sacchariflorus, sinensis, sacc/Robustus, sacc/lutarioriparius, and floridulus (Figure 4.11). The possible reason for the difference between this study and the published literature could be due to the number of *M. sacchariflorus* present in this study. In Robson *et al*.'s (2012) study they did not measure stem counts, instead they used a transect count. This is an estimation of stem count and not the actual number of stems as used in this study.

Very few traits show relation to yield in the hybrids. In the sacc/sin hybrids canopy height was demonstrated to be the most influential trait for yield.

Canopy Height

Canopy height had the highest variance explained in all the models. This implies that this trait is predominantly genetically driven, with only small amounts of variance attributable to environment or error. Many SNP markers detected in this study can be mapped to known QTL that affect canopy height in *Sorghum* (Lin *et al.,* 1995; Ritter *et al.,* 2008; Shiringani *et al*., 2010).

Stem Thickness

For stem thickness a high fraction of the variance was explained in the *M.*

sacchariflorus subset, but was much less influential in the *M. sinensis*. There were very few studies published at the time of writing in *Sorghum* for stem diameter. The regions detected by the random forest model lie on the same chromosome as was detected in QTL studies of stem thickness (Zou *et al*., 2012).

Base Diameter

Base diameter could not be explained by the modelling in this study. Growth that alters base diameter occur under ground. Therefore it is likely that what is measured above ground many not reflect the rhizome spread. In later years, when plants are more established below ground, it is possible that the base diameter may become more representative of the actual plant morphology. Base diameter in this study may be controlled by the size and quality of rhizome used at planting; however, this was not quantified and therefore cannot be included in this study.

Stem Counts

Stems counts appear to have a degree of genetic control with almost 50% of the variance explained by markers in *M. sacchariflorus* subset. Stem count was also related to yield and this differs from published literature (Robson *et al.,* 2012). Several of the SNP markers detected matched with known Sorghum QTL (Paterson *et al.,* 1995; Hart *et al.,* 2001; Shiringani *et al.,* 2010).

Tallest Stem

Tallest stem is often highly correlated with canopy height. However, the correlation breaks down when a plant transitions to flowering (Figure 4.4). Within the *M. sacchariflorus* subset, the genomic regions that relate to both canopy height and tallest stem are similar. This is due to the correlation between the traits, and will not be

confounded by flowering because of the low flowering rate within the *M. sacchariflorus* genotypes. In the *M. sinensis* subset the tallest stem has a complex relationship to canopy height depending upon its flowering stage (Figure 4.4). Different regions were identified between the *M. sinensis* tallest stem and canopy height.

Yield

Linking yield to genetics directly would allow for MAS or GS breeding for high yielding varieties. Yield is a complex trait which is made up of many phenotypical traits (Robson *et al*., 2013). This study has demonstrated that there are markers that can be used to associate genotype to high and low yielding plants. Many of the markers which were detected for yield appear to be over dominant. Also several loci were identified where the high yielding genotypes had no allele.

Over dominance appeared to play a role in increasing yield in *Miscanthus,* within both genotypic subsets markers were identified where the heterozygote allele associates with higher yield. Over dominance is were a heterozygote has a stronger phenotype than the homozygotes, in this case the plant produces a greater amount of biomass. As *Miscanthus* is also most exclusively an out breeding plant, which means that the number of heterozygotes is likely to be high, and we known that the hybrids such as *M x giganteus* are high biomass producers. This could potentially suggest that over dominance may have a strong role to play in Miscanthus breeding. This has been observed in other plant species where over dominance and epistatic interactions were major factors for heterosis in yields (Li *et al*. 2001; Semel *et al.,* 2006).

Moisture Content

The random forest approach was unable to fit a model to moisture content, with the result explaining almost none of the variance. All the traits that have thus far been

measured in the GBS study were taken on immature plants. It could be that moisture content might become more dominated by genetics but more measurements would be needed to confirm this.

It has been reported that very few QTLs for moisture content could be detected in *Sorghum* (Felderhoff *et al*., 2012). This might suggest that moisture content maybe potentially difficult trait to breed for. This would agree with the results from the model in which very little variance could be explained by the marker analysis results.

4.5 Concluding Remarks

From the results of this study it would appear that random forest is capable of detecting features in high marker number data sets to generate associations with phenotypes. However random forest will firstly select markers that explain strong genetic differences between genotypes, such as those that explain species variations. This still correctly identifies markers that associate with the traits. However without separating out the two species it would be difficult to know which markers related to which species without investigation the tree structures within the random forest model. To counteract this genetic dissimilarity was used to first separate the species. This approach is in someway similar to the methods used in GWAS when calculating kinship, however this is not required for the random forest analysis to work, only to improve readability. In GWAS kinship is an integral part of the analysis, it is often calculated as a percentage change of two genotypes sharing the same allele, whereas the method used in the random forest analysis was a simple scoring system which quantified how dissimilar two genotype are.

Miscanthus has high levels of synteny with *Sorghum*, therefore comparisons of marker loci can be performed between the two species. It was shown that the markers

detected by the random forest approach detects markers in regions of the *Sorghum* genome known to contain QTLs. This implies that several QTLs from *Sorghum* are in common with *Miscanthu*s. It also demonstrates the ability of the random forest method to efficiently detect markers in the regions that associate with traits.

This study demonstrates the effectiveness of importance score mapping via random forest analysis. Only the highest importance markers have been looked at and there are many others that show less importance, but may be significant. It has been suggested in the literature that traits are unlikely to comprise many large QTL, and are more likely to comprise a few large effect QTL and many smaller effect QTL(Buckler *et al*., 2009). Therefore further investigation of the lower importance markers may reveal many more important associations.

The data analysed in this study only looked at first year observations where the plant is still immature. Many studies suggested that using multiple year data in genetic analysis is necessary in order to account for environmental effects (Ritter *et al.,* 2008; Gifford *et al.,* 2014). Some of these traits may have a greater variance explained in a mature plant. However to know this, more observations need to be performed on the same trial. This will allow models of the relationship between genetics, phenotype and maturity to be developed. Potentially different genetic regions could control phenotypes in maturity. Alternatively the same regions may control a trait in mature and immature *Miscanthus* but with a different magnitude. Immature phenotypes may also be confounded by the affect of the amount of rhizome used in planting.

The highest yields were seen in the hybrid species, which does imply that hybridization may be a key method in the development of high yielding genotypes. *M. sacchariflorus* tended to be taller plants consisting of fewer, but thicker stems, whereas the

M. sinensis tended to be shorter with more and thinner stems. A high stem count with large stem numbers and relatively thick stems are all traits that are found the naturally occurring hybrid *M. x giganteus*.

5 Exploiting Machine Learning in Modelling of Genotype-by-Phenotype-by-Environment (GxPxE) Association

5.1 Introduction

Several scientific questions have been intensively discussed when studying genetic and phenotypic interactions: 1) what is the effect of the environment on genotype and phenotype? 2) how does the environment change the way the genome behaves? and 3) how to link these effects with the phenotype?

Changes in the relative performance of genotype across different environments are referred to as genotype-by-environment interaction. A major objective in plant breeding programmes is to understand this interaction and assess the suitability of individual crop genotypes for agriculture purposes across a range of environmental conditions.

The classical formula of genotype-by-environment interaction is defined as:

$P=G+E$ (Falconer & Mackay, 1996)

Where P is the phenotype and is a combined function of genetic effects (G) plus the environmental effects (E). The genetic variable can be broken down further.

$G=G_A+G_D+G_I$ (Falconer & Mackay, 1996)

Where A, D and I are additive, dominance and interactive or epistatic genetic effects. Additive (A) is the simplest variation in which genes/markers are quantified as a gain or reduction in a phenotype. Dominance (D) is the interaction effects between alleles. Interactive (I) or epistatic effects are more complicated. They can be in the form of interactions between different genes or can change DNA structure via methods such as

methylation, which subsequently modify the observed phenotypes.

Methods associating genotype to phenotype generally focus on additive interactions as their effects can be easily quantified by observing phenotypes. Responses to the environment are likely to be linked with epistatic interactions. This includes gene interactions, such as genes that respond to changes in the environment and subsequently lead to new pathways being expressed (Shinozaki *et al.,* 2003).

Previous studies have attempted to use ANOVA (analysis of variance) (Lukens & Doebley, 1999) and mixed modelling (Wang *et al.,* 1999; Ungerer *et al.,* 2003) to find the link between known QTL with epistatic effects which are commonly associated with environmental interactions. These researches mainly focused on the first equation ($P=G+E$) where all genetic effects are considered as one single variable.

Meteorological effects are complex. Many factors, wind, rain, sun, cloud cover, minimum and maximum temperature, are all linked to each other and have combinational effect on the plant. Any genotype grown in a field could experience a wide range of conditions and stresses. These could include exposure to different soil types, fertility levels, moisture contents, temperatures, photoperiods, biotic and abiotic stresses.

Drought is another stress that plants often experience. Radiation from the sun, amount of wind and temperature all contribute to the loss of water from the soil. Irrigation can be used to top up the rainfall; however irrigation comes with high financial cost. On the other hand, water cannot be reduced when extreme weather events lead to flooding. Therefore, it is essential to understand how the genetics and environment work together to regulate the plant phenotype and its physiological responses to water deficit (Ings *et al.,* 2013).

Several studies have attempted to link phenotype with genotype (Snape *et al.,* 1977; Semel *et al.,* 2006; Neumann *et al.,* 2010). They mainly concentrated on single location and single year observations as a way to simplify the environmental variation (Smith *et al.,* 2005). Environmental variation can occur due to water limitation (Warrick & Gardner, 1983) or nutrient availability (Jin & Jiang, 2002). Small spatial variations, such as local access to nutrients, even within the same field trial have been shown to have effect on a plant's phenotype (Trangmar *et al.*, 1987). To address this environmental variation, randomised blocks plot designs are normally used and then analysed using a spatial model.

Genotype by environment interaction has significant influence on the efficiency of crop improvement. An understanding of the genotype stability across environments can help in the determination of their suitability for the fluctuations in growing conditions that are likely to be encountered. Environmental interactions can potentially have strong effects on a plant’s phenotype (Sultan, 2000). Environmental factors could have a much greater effect on perennial species than on the annual crop. Several studies have looked at modelling *Miscanthus* yields in relation to the environment, although these are often limited to small number of genotypes (Hastings *et al*. 2009a; Pogson 2011).

Miscanthus is a perennial grass and is known to produce continuous yield for at least a decade (Gauder *et al*. 2012). Any change in climate could impact on plant health and lead to a reduced yield. Therefore any new variety needs to be able to adapt to the changing environments. It is vitally important to better understand the interactions between the genotype and environment.

Machine learning is known for its ability to detect complex patterns in high dimensional data. Learning methods (Hastie *et al.,* 2009) are capable of modelling complex interactions by “learning” how a set of inputs leads to an output. Machine learning

could provide an alternative to model these interactions. By performing attribute selection (Kononenko & Hong, 1997), machine learning algorithms can remove the measurements which do not effect the response. It can reduce the complexity of the resulting model, and informing breeders which environmental factors actually affect a given trait.

Drought is one of the major concerns for modern crop breeding programmes. Climate models have predicted that future water availability will become sparse due to climate change (Schröter *et al.,* 2005; Olesen & Bindi, 2002). There is a consensus that food crops are more important and should be given priority to the use of water resources. It means that water use will have to be limited when growing the non-food crops such as bio-energy crops (Pimentel *et al.,* 2008).

This study aims to improve the understanding of how environment factors, in particular the drought effect, influence the change in flowering time by applying the machine learning approach to improve the modelling efficiency. The *Miscanthus* flowering time data collected from the 2TT population has been used to build this GxE interactions model to investigate how water availability can affect the flowering time in *Miscanthus*. It will also be used to investigate the effects of future climate predictions on flowering time in 2020.

5.2 Case Study – Modelling of Environmental Effect on Miscanthus performance using Machine Learning

Wild *Miscanthus* grows naturally in a wide range of latitudes ranging from 45° to 18° with great diverse climate conditions. The 2TT population consists of several hybrids, horticultural collections and wild accessions. Table 5.1 shows a summary of the climate conditions under which the wild germplasm, featured in this study, were collected.

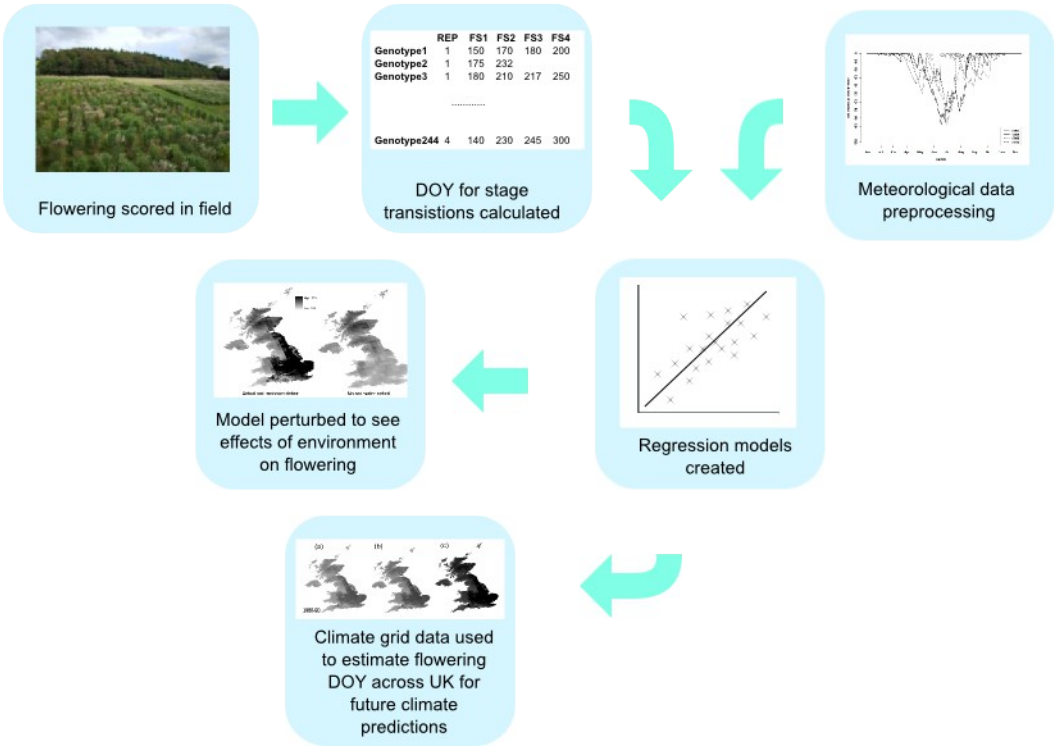

Figure 5.1: The analysis pipeline used in this study. Data was collected from the 2TT trial. Data was processed to obtain a DOY value for flowering stage transition. A regression model was established to investigate the effects of soil moisture deficit and predict effect of changing in climate condition on flowering time in Miscanthus.

However for the early horticultural collections the site of the original collections is unknown so the climate data cannot be included. Therefore the table only represents a subsample of the population.

Measurement	**Minimum**	**Maximum**	**Mean**
Annual Rainfall (mm)	493	2848	1403.7
Minimum Temp (ºC)	-24.6	5.6	-4
Maximum Temp (ºC)	18.2	27.3	23.3
Annual Degree Days	143.6	1800	1318

Table 5.1: Summary of climate data from collection sites of the Miscanthus germplasm in 2TT trial.

It is believed that *Miscanthus* flowering time contributes to the yield. Previous work have shown late flowering *M. sacchariflorus* can lead to a higher yield (Jensen *et al*.,

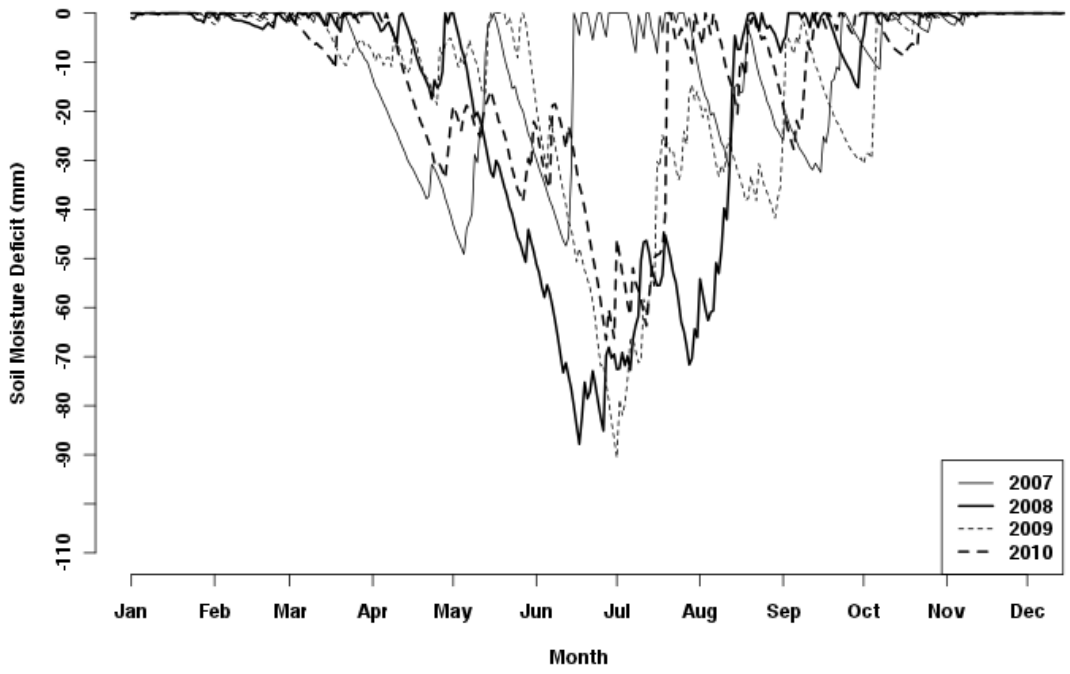

Figure 5.2: The soil moisture deficit for the trial is shown. This was predicted for each year using data from a near by meteorological station and the Penman-Monteith equation.

2013). This study also showed that *M. sacchariflorus* flowering requires a short day length to initiate. On the other hand, *M. sinensis is considered to be* day length neutral for flowering (Deuter, 2000).

Observations of flowering time used in this study were taken over four years and meteorological data was retrieved from a nearby meteorological station. The machine learning approach was applied to build a model to study the effect of the temperature and water availability in the soil.

5.3 Results and Discussions

Figure 5.1 displayed the analysis workflow used in this study. Flowering time and meteorological data were collected and pre-processed. The data used come from a

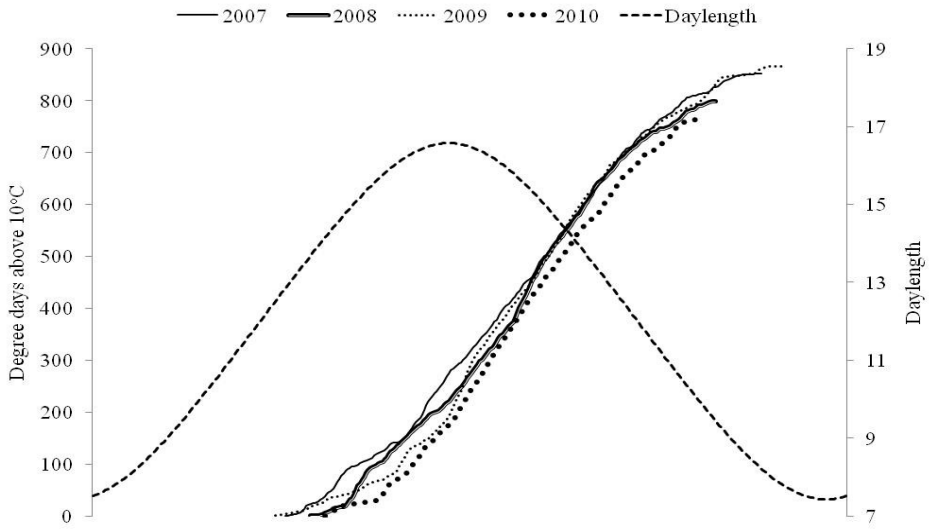

Figure 5.3: The cumulative degree days values for the four years in which flowering observations were taken. Also the day length profile for the trial location was illustrated in this figure. The degree days of each year is shown to be very similar, reaching a maximum of approximately 900. Which is half of the maximum collection degree days.

previously published study (Jensen *et al.,* 2011a). Four flowering stages were measured on 244 genotypes in this study; flag leaf emergence (FS1), panicle emergence (FS2), 50% stem flowering (FS3) and 80% stem flowering (FS4). Each week the flowering stage was recorded for each genotype, from this data the day of year (DOY) for each stage initiation was calculated. This was done by looking for the first time each stage was observed. A regression model was created using this dataset to predict day of year (DOY) values for two flowering stages (FS1 and FS2) of *Miscanthus.* The model was then used to investigate the effects of soil moisture deficit on flowering in *Miscanthus* to predict the impact of changing climate conditions on flowering DOY. The attributes used in this model were degree days and soil moisture deficit.

5.3.1 Meteorological Data Preprocessing

Soil moisture deficit (SMD) was calculated based on a close bucket model using the FAO standard for Penman-Monteith (Allen *et al*., 1998). The trial is considered as a closed

system where the only water source is coming from rainfall. It was assumed that there will be no water run off and that the only water leaving the system is through evapotranspiration. The field was always assumed to be saturated on the first day of the year, therefore SMD was set to 0 on day 1. The value of SMD is calculated for the year seen in Figure 5.2.

Unsurprisingly the highest SMD was reached during the summer when rainfall is low and temperatures are high meaning greater water loss without the rainfall to recover the soil moisture level. The highest SMD used in this study was approximately 90mm. The wilting point varies depending upon field type, one study suggested that for a clay like soil that its field capacity, its ability to hold water, would be 120mm, resulting in a wilt point of 60mm for *Miscanthus*. (Hastings *et al.,* 2009a). Wilt point is the highest SMD value from which a plant can recover. If the SMD becomes higher than the wilt point a plant will not be able to recover its turgidity. The wilt point of this trial was estimated to be 150mm with a field capacity of 350mm (Jensen *et al.,* 2011b). The plants within the 2TT trial would have experienced drought stress without reaching the wilt point, therefore all the plants should have recovered from any drought effects.

Degree days were calculated, to be utilised in the model in order for temperature effects to be accounted for, over four years with a base temperature of 10°C. Figure 5.3 is the calculated degree days values for 2TT trial. The highest degree days during the four year period were 900. As seen in Table 5.1 the average degree days of germplasm collection sites is around 400 degree days more than the plants experienced at the trial site.

The attributes for the linear regression model were degree days and SMD. Degree days were created by taking the cumulative degree days for 7 days. SMD values were

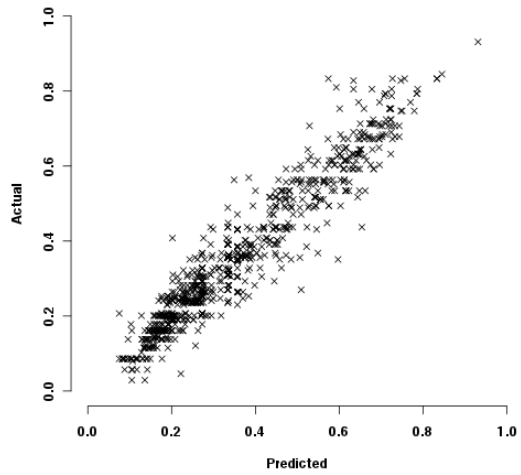

Figure 5.4: Normalised observed DOY versus the predicted DOY of flowering stage 1 FS1 using SMD and degree days as attributes in linear regression model.

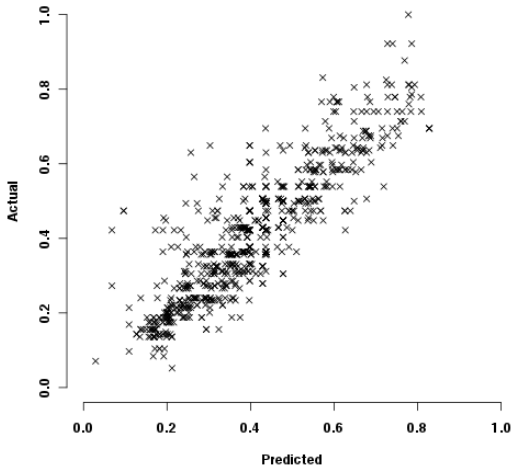

Figure 5.5: Normalised observed DOY versus the predicted DOY of flowering stage 2 FS2 using SMD and degree days as attributes in linear regression model.

converted into a ratio using the following formula, this makes the measurement of SMD dimensionless.

$$\frac{Actual\ Soil\ Mositure\ on\ Day\left(d\right)-Wilt\ Point}{Field\ Capacity-Wilt\ Point}$$

The average ratio over 7 days was generated as one of the attributes. They were calculated over the whole growing season between the first and last frost. The whole season was on average 33 weeks. This made 33 attributes for degree days and SMD ratio. A total of 66 meteorological attributes were included in the model. The model did not include photoperiod since the data were only collected at one site.

5.3.2 Regression Analysis

Each genotype was programmed as a factor variable with its own accession number. It was combined with the meteorological data to create an input matrix. The DOY score was calculated and model was fitted to provide the prediction for each stage of flowering.

Method (Weka)	Data Set	Mean Absolute Error	Prediction Correlation
Least Median Squares	FS1	0.0732	0.7361
Artificial Neural Network	FS1	0.1192	0.6697
Decision Tree (REPtree)	FS1	0.1044	0.6384
Decision Tree (M5)	FS1	0.1441	0.1321
Linear Regression (M5 Parameter selection)	FS1	0.0536	0.89

Table 5.2: Summary table of the machine learning methods attempted for modelling Miscanthus flowering time

Each stage was considered to be independent of the previous stage.

The predictive model was then created using the machine learning programming package WEKA (Hall *et al*., 2009). A subset of data was held back in reserve as a validation data set for later testing of the model.

Several different machine learning algorithms were tested during this study in order to identify the model with the lowest error and best fit (Table 5.2). Decision trees, artificial neural networks, and linear regression with M5 attribute selection (Hall *et al.,* 2009) were tested. Out of all the algorithms tested, the simplest and most accurate model was linear regression with M5 attribute selection. It can remove attributes by stepping through the problem domain and removes the one with the smallest coefficient until no improvement is found in the error estimate using the Akaike information criterion for model selection. The model was tested using 30-fold cross validation. Linear regression had the lowest error and highest correlation of all the algorithms tested. Once a model was created, it was validated against the validation data set to ensure overfitting had not occurred.

All the models established were unable to predict the following two stages of flowering, FS3 and FS4, with the error being much larger than the observation frequency.

FS1 and FS2 models had correlation coefficients of 0.89. The FS1 model had a normalised mean absolute error of 0.0536; and the FS2 model had a normalised mean

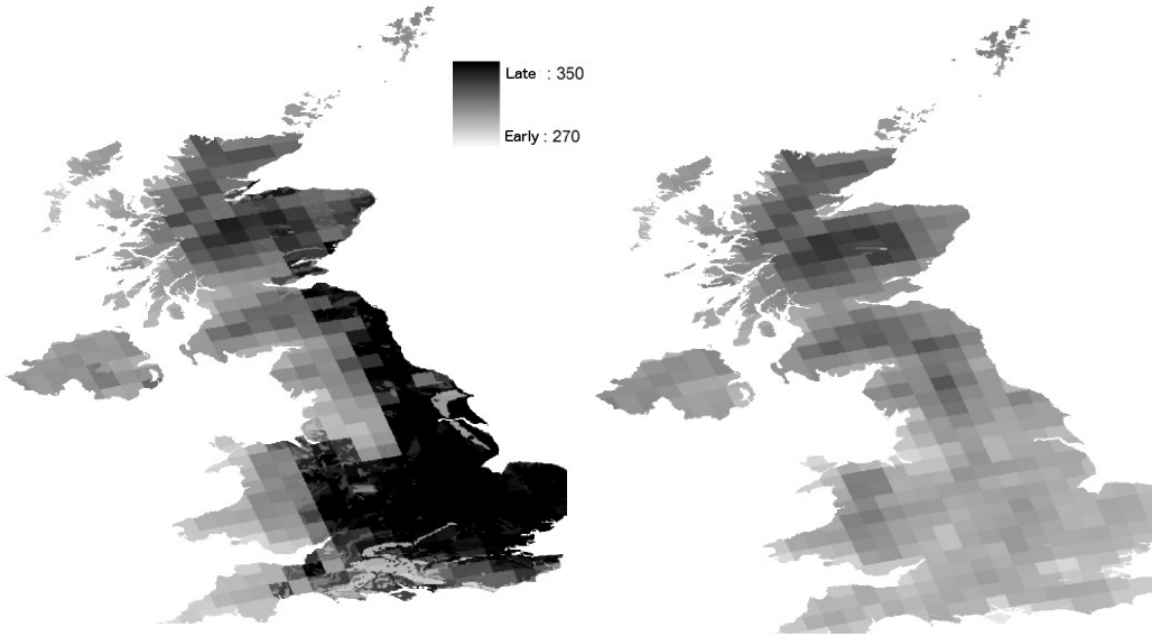

Figure 5.6: The graph on the left presents the expected DOY with soil moisture deficit and the graph on the right presents the expected DOY without soil moisture deficit. Earlier flowering has been observed without deficit in soil moisture.

absolute error of 0.0565. These models were then tested against the validation datasets set aside earlier. The results on the validation data are as follows; the FS1 model gave a correlation of 0.956 and a normalised error of 0.0382 (Figure 5.4). The FS2 model gave a correlation of 0.884 and an error of 0.0581 (Figure 5.5).

5.3.3 Predicting *Miscanthus* flowering under different climatic conditions

After the models were developed a subset of *M. sinensis* genotypes were selected for further investigation. The resulting formula from the linear regression model was further used to model flowering DOY across the UK using the UKCP09 meteorological data sets on a 25km grid (Jenkins et al., 2009) for a subset of genotypes in 2TT (Jensen *et al.,* 2011b). The data was visualised using ArcGIS (ESRI 2011).

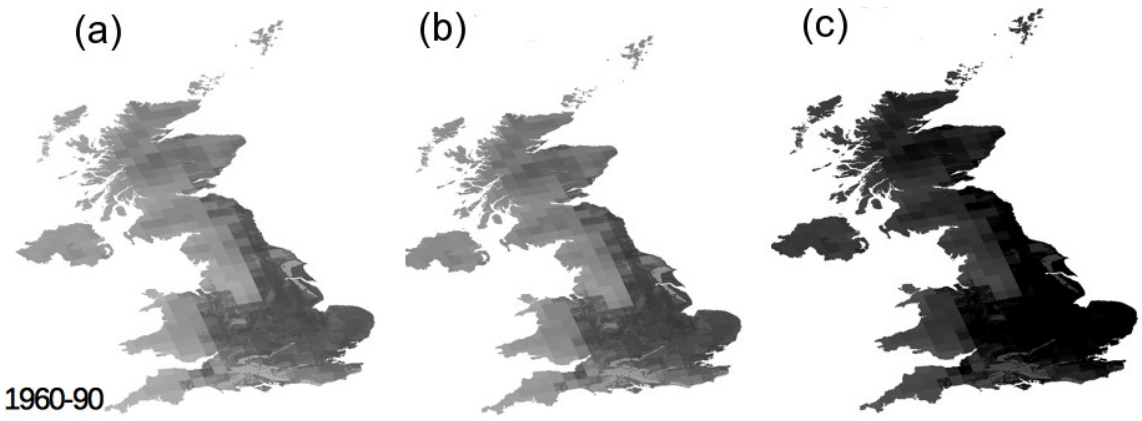

Figure 5.7: Flowering time predictions for three modelled genotypes created using climate data from 1960-1990. Three genotypes are a) early, b) mid-season, c) late flowering.

Figure 5.8: The resulting 3 modelled genotypes based on the climate data predicted for 2020.

Effects of SMD on *Miscanthus* Flowering

The predicted DOY value for a late flowering *M. sinensis* genotypes across UK shows variation of flowering under normal conditions and without drought (Figure 5.6). The map on the left shows the expected flowering DOY for this particular genotype under the effect of soil moisture deficit (drought). The right hand side map shows the expected flowering DOY with irrigation, so no SMD was experienced.

Many *Miscanthus* accessions enter senescence and do not flower (Jensen *et al.,*

2011a). Therefore, in order to avoid the DOY value runs over 365, a maximum DOY was set to be 350 for those non-flowering genotypes when visualising results.

The map on the left hand side demonstrated that when SMD factor was included in the model, there is a shift towards late flowering (Figure 5.6). When no drought (zero SMD) was present, the same genotype is likely to flower earlier at any location throughout the UK. Although slightly later flowering was observed in Scotland, which was an effect attributed to lower temperatures with less degree days. We can conclude that, in general, the introduction of drought effect causes the flowering time to be delayed. The effect is much greater in the east whereas the far west of the UK remains mostly unchanged. This is likely caused by the higher rainfall seen on the west coast when compared to the east. The model also suggests that for this particular genotype it may not flower if grown on the eastern side of the UK. It highlights the great variation in flowering time even within the same country due to the difference in soil type and water availability.

Miscanthus Flowering Under Future Climates

This study selected three *M. sinensis* genotypes from the 2TT population to predict the flowering time based on climate data of 1960-1990 and the predicted future climate data of hi scenario 2020 for this investigation (Jenkins et al., 2009). Only *M. sinensis* were used in this investigation, as *M. sinensis* has been shown to be day length neutral (Deuter, 2000), to eliminate the day length variable. The three genotypes include one early flowering (a), one flowering in the middle of season (b) and one late flowering (c) genotype.

Two climate data sets were selected. One is the historical climatic dataset between 1960-90 and another is the predicted climatic conditions for 2020. Each genotype was

modelled under these two climatic scenarios. The results of 1960-90 models are illustrated in Figure 5.7 and the 2020 predictions are displayed as Figure 5.8.

From the predictions maps of 2020, all three *M. sinensis* genotypes exhibit earlier flowering under future climate predictions of 2020. Though the pattern is still the same, there existed a difference in flowering time between the west and east of the UK which is most likely due to the differences in rainfall. The late flowering genotype (c) shows the largest shift to flower late. Under the predicted future climate condition, the late flowering genotype will flower in many more locations. Many climate studies have predicted great increase in drought condition (Schröter *et al.*, 2005; Olesen & Bindi, 2002). Therefore, one might expect flowering in the 2020 models to be late flowering as was seen in Figure 5.6 when drought was included in the model. However future climate models also suggest an increase in temperature (Hansen *et al.,* 2006). The hi scenario predictions (Jenkins et al., 2009) utilised in the model predicts the following changes to the climate, firstly they suggest that there will be an increase in daily mean temperature, both minimum and maximum by an average of 2C. Secondly that precipitation will be down in the summer by an average of 40%. It is likely that the increased temperatures are more influential on flowering time than the drought effects on this particular genotype. and lead to the shift towards an earlier flowering (Figure 5.8).

5.3.4 Discussion

The predicted results have revealed two important effects of climatic variation on *Miscanthus* flowering time. Firstly SMD, and therefore drought, causes *Miscanthus* to delay its flowering time. Secondly the increasing temperatures associated with future climate predictions will lead to early flowering in *Miscanthus.*

Drought effects on *Miscanthus* Flowering

This study demonstrated that drought has a delaying effect on *Miscanthus* flowering time (Figure 5.6). For some annual crops, drought will lead to earlier flowering (Franks *et al.*, 2007; Heschel & Riginos, 2005); however in *Miscanthus*, drought seemed to delay the flowering. Since *Miscanthus* is a perennial crop it may be a favourable strategy not to expend resources on flowering when stressed but instead enter into senescence. *M. x giganteus* and *M. sacchariflorus* have both been shown to enter senescence when under drought stress (Clifton-Brown, 2000). This indicates that more resources are reserved for the following year in which the flowering conditions may improve. However this phenomenon was not found in the *M. sinensis* hybrid investigated in this study.

It appeared that drought causes *Miscanthus* to flower later for the *M. sinensis* genotype investigated. This may be advantageous as studies have suggested that late flowering leads to increased yield (Gonza *et al.,* 2001; Jensen *et al.,* 2013). However creating drought situation is much more difficult than preventing it. Also drought could potentially decrease yield (Lewandowski & Heinz, 2003).

Effects of Climate change on *Miscanthus* Flowering

It has been suggested in several studies that there is a link between flowering and yield in *Miscanthus* (Gonza *et al.,* 2001; Jensen *et al.,* 2013). Using the future climate models the predicted flowering was shown to initiate earlier in *Miscanthus* (Figure 5.8) This could potentially result in a diminished yield for biomass production in the UK.

Machine learning and its role in GxExP Modelling

The models created in this study allowed for further development of two hypotheses on how environmental factors affect *Miscanthus*' flowering time. Firstly drought can delay

flowering. Secondly the effects of climate change will lead to earlier flowering. Attribute selection using the M5 method can make sure that only influential attributes are included in the model. It can therefore build high confidence on the effects of both degree days (temperature) and SMD (drought) on flowering time. As any non-influential attributes would have been removed and therefore would have no effect on the flowering prediction. The use of cross validation and a validation data set meant that the model established in this study was not overfitted to the training set.

Improving Genetic Component of Modelling

High density SNP markers are being generated on several mapping population and trials at IBERS and will be available in the near future. By introducing genetic markers into the analysis, a more comprehensive model of GxExP could be developed. This model could then be used to predict new germplasm, as the effects would be attributed to markers rather than a genotype. However a more effective machine learning algorithm may be needed to handle those complex datasets.

Model Improvements

Although the model created by machine learning is effective at predicting flowering time from the observations, there are several potential improvements that could be made if a similar study is performed again.

Firstly the frequency of observations was low either being weekly or fortnightly in the current study. Weekly or twice every week would potentially lead to more accurate observations and therefore potentially reduce the error of the model.

Secondly a model of senescence should also be developed in conjunction with the flowering model. This could enhance the model so that it could label particular genotypes

as non-flowering for a given environment, rather than a cut-off point being applied as was used in this study.

Another area this model could be improved is by the selection of a more informed SMD model. The SMD model used in this study is over simplistic due to the use of the closed bucket assumption. Firstly it assumed that all the available water is evenly distributed across the trial. In reality it is more likely that pockets of higher or lower water concentration are found across the field. Secondly, it was assumed that the fields water was a closed system and that water only came into the system through rainfall. Thirdly, another assumption was that water run-off from the field did not occur during rainfall. Given that this trial is situated on a slope, it seems likely that a run-off did occur. To overcome this issue of water being unevenly distributed, a direct water model should be developed by using reflectometers to measure the moisture directly at several locations within the trial to create a spatial model. Alternatively a more informed model of water availability should be developed which accounts for slope and potential run-off.

Temperature changes were parameterised using the degree days formula. Degree days is often used in modelling crop interactions. It represents a linear transformation of temperature into a cumulative measurement of units of degrees above a given base per day. This minimizes the effects of extreme temperature fluctuations as it discounted cold temperatures where many low temperature days will have a degree day value of 0. However, a late frost occuring after emergence could have an effect on new growth, causing delay on the plant development as existing shoots may have died and the new ones will take time to recover from the lost growth. Therefore, another possible improvement to the model will be using the minimum and maximum temperature to allow a model to understand the range experienced by a plant. Also the base temperature of

Miscanthus has never been officially established and is only assumed to be 10°C, but again this is another simplification of a plant environmental interaction.

One final improvement that could be added to this type of modelling is the use of multi-location trials. Within this study, variations between the years were very small. This was found in the soil moisture that the amount of water loss is similar between years with the lowest value around 90mm. The inclusion of additional sites which experience more or less drought would enhance the model better to understand the drought effects on flowering time in *Miscanthus*.

5.4 Concluding Remarks

The objective of this study is to demonstrate that the machine learning approach can be an effective tool in establishing GxE interaction models for better understanding of the environmental effects on a given trait. The established model was used to investigate the consequences of SMD deficit, a measure of drought, on flowering time in several *Miscanthus* genotypes. Due to the application of attribute selection, this study showed that both SMD (drought effect) and degree days (temperature) are important in the control of *Miscanthus* flowering time. Day length was shown in previous studies to have strong effect on *M. sacchariflorus* flowering time (Jensen *et al.,* 2013). However as the data were only collected at a single site with little variation in observations, further study is needed.

The model developed allows breeders to understand the likely effects of drought and increasing temperature on several *Miscanthus* genotypes. Based on the knowledge from both published literature (Jensen *et al.,* 2013) and the prediction results of this investigation (Figure 5.8), we can conclude that the yield of Miscanthus may be significantly diminished due to climate change over the course of their life time for certain

genotypes. Flowering has been shown to be highly heritable (Slavov *et al.,* 2013). Breeding late or non-flowering genotypes is therefore a preferable goal to maintaining crop viability. Based on the future climate prediction, it is also necessary to develop genotypes that are more tolerant to drought and high temperatures for Miscanthus to eliminate the unwanted effect of climate change.

The understanding of drought delaying flowering time (Figure 5.6) can be used to inform the breeder when making new crosses. For example, by preventing drought on one parent through irrigation, breeders could achieve flowering synchronisation among different genotypes for crossing.

Future climate models suggest more volatile weather patterns and a higher frequency of extreme weather events. Understanding the effects of GxE interaction in particular for perennial crops is therefore of high importance in crop development.

It is evident from this study, the application of machine learning to modelling allows for better understanding of the complex interaction between GxE. Furthermore with the inclusion of genetic data in the future, models could be created to predict phenotypes for a given environment using a given marker profile. Modelling results could be incorporated into a MAS programme, allowing the breeder to select the optimised traits to suit different environments. Understanding the genetics of additive, dominance and interaction effects is still in its infancy. Machine learning, as demonstrated in this study, can be an effective approach to model environmental effects. Combined with genetic data e.g. QTL data (Gifford *et al*. 2014), enhanced models could be created to estimate the effects of dominance and interactions from genetic variance.

6 Discussion and Future Research

6.1 Discussion

The main objective of this research is to apply machine learning approach to model trait marker associations and environment interactions. Throughout this thesis machine learning has been described how it can underpin the QTL discovery, analysis of high throughput markers and the understanding of relationship between phenotype, environment and genotype.

6.1.1 Machine Learning for Genetics Research

In general, this research contains 3 studies for the energy crop *Miscanthus*. Machine learning was utilised throughout this thesis, leading to the development of several machine learning based methods: a machine-learning based QTL analysis tool (RFQTL), perform analysis of genotyping-by-sequencing (GBS) data and build a computational model to study the GxExP interactions.

Machine learning based QTL analysis tool (RFQTL)

Random forest (Breiman, 2011a), a machine learning algorithm was used to develop a QTL analysis tool, referred to as RFQTL, to identify high importance markers associated with QTLs. A flowering time mapping family of 236 *M. sinensis* (Ma *et al.,* 2012b) was genotyped and the data generated was used to perform QTL analysis using RFQTL. The analysis results were compared to conventional QTL analysis methods, MapQTL (Ooijen, 2004) and GoldenHelix (Golden Helix, Inc., Bozeman, MT, www.goldenhelix.com).

Two stages of flowering time observations were taken, flag leaf emergence and

panicle emergence (2009 – 2011). All three QTL analysis methods detected a high importance QTL on LG04 for both stages as described in Chapter 3 (Figure 3.3, Figure 3.4, Figure 3.6, Figure 3.7, Figure 3.8 and Figure 3.9). Comparisons to *Sorghum QTL study* also revealed a homolog QTL between *Miscanthus* and *Sorghum* on flowering time (Lin *et al.,* 1995).

Year and age were the highest importance attributes found in the model. Therefore attributes to account for variance between years were added to build the basic model. Attributes of photo-synthetically active radiation, temperature (minimum and maximum) and rainfall were also used to improve the model with two more QTL on LG06 and LG11 were found.

Only the panicle emergence dataset was used for the 2013 analysis with a higher frequency of phenotyping. Similar results with 2009-2012 analysis were generated. Flowering in 2013 was delayed and this is most likely due to high temperature and low rainfall, which could account for the new QTL.

RFQTL has shown to be able to produce consistent results with conventional QTL analysis methods. RFQTL is a more computationally efficient and is not limited to linear relationships which only function well with additive genetic variance like conventional QTL methods. Non-linear models have the potential to detect epistatic interactions. It is evident from this study that RFQTL is a powerful method for QTL analysis. It provides another tool in the arsenal for the molecular breeder to economically detect QTL.

Machine Learning for GBS data

Genotyping by sequencing (GBS) is an economical technique for the development of large numbers of genetic markers in complex genomes (Elshire *et al.,* 2011). Machine

learning was applied in this study to perform marker trait association using the random forest algorithm as described in chapter 4.

244 genotypes were selected from the germplasm collection at IBERS and GBS was performed on 179 accessions with 3778 SNP markers were generated. Genetic dissimilarity was measured between the genotypes and it revealed three distinct groupings of the 179 accessions. The groups, *M. sinensis* and *M. sacchariflorus*, were analysed for marker traits associations. Seven traits were observed and modelled using random forest; and the variance explained was summarised in Table 6.1. Stem count had the strongest correlation to yield. However, this differs from the published studies where canopy height has the highest correlation with yield (Robson *et al.,* 2013). This may be caused by the use of different phenotypical measurement methodologies in the studies.

Trait	**Var. Exp. *M. sinensis* subset**	**Var. Exp. *M. sacchariflorus* subset**
Canopy Height	43.03%	66.97%
Tallest Stem	33.06%	64%
Stem Diameter	22.5%	56.51%
Stem Count	26.06%	49.99%
Base Diameter	Unable to fit model	Unable to fit model
Dry Weight	13.62%	34.11%
Moisture Content	Unable to fit model	Unable to fit model

Table 6.1: Variance explained by the random forest model for each trait and species group

Within this study homologous QTL within *Sorghum* were also investigated with many common QTLs detected (Lin *et al.,* 1995; Paterson *et al.,* 1995; Hart *et al.,* 2001; Ritter *et al.,* 2008; Shiringani *et al.,* 2010; Zou *et al*., 2012). Markers detected responsible for yield were shown to display overdominance and this finding has confirmed previous studies that overdominance effect has displayed high correlation to yield in other crop species such as tomato (Semel *et al.,* 2006).

Based on the results of this research we have demonstrated that machine learning has the capability to analyse massive markers generated from high throughput technology and associate the complex relationship between markers and traits of interest. Unlike conventional statistically based methods, the random forest algorithm was able to perform a much more dynamic analysis from different perspectives. Consequently, this allows new scientific insight to be uncovered and novel hypothesis to be formulated from high dimensional data sets.

Machine Learning for GxExP Association

The relationship between genotype, environment and phenotype are complex in nature with many attributes may be hard to obtain with accuracy. In chapter 5 machine learning was used to model the effects of the environment on flowering time in *Miscanthus* using data from a previously published study (Jensen *et al.,* 2011a). Meteorological data consisting of degree days and soil moisture deficit ratio observed over the growing season were included in the model to study and predict the environmental effects on flowering time.

Through the application of data mining, large volumes of meteorological data can be pre-processed and used to identify which attributes have strong effects on the trait being investigated. This simplifies the model and produce better results than the conventional process models that require large amounts of complex observations to investigate the relationship between meteorological factors and phenotypic or physiological response (Davey *et al., In preparation)*.

A germplasm collection consisting of 244 accessions was used to develop a model of the effects of temperature and drought on flowering time. The developed model have

suggested that soil moisture deficit (SMD) does delay flowering time in the *M. sinensis* genotypes which were selected for further investigation as shown in the conclusion of Chapter 5 (Figure 5.6). The results also reveal that based on future climate scenarios, flowering will happen earlier for all the genotypes studied in this research (Figure 5.7, Figure 5.8).

6.1.2 Advantages of Machine Learning in Genetic Studies

Machine learning has many advantages over the conventional statistics methods when associating genetics with environment. Many conventional methods are linear and model only additive genetic variances. It has been suggested that non-linear methods may be able to detect epistatic effects (Jannink *et al.,* 2010). The analysis using random forest algorithm was able to detect markers that display epistatic relationships. The first evidence can be found in the rice analysis described in Chapter 2. A SNP marker was detected in this study, which was previously undocumented and did not appear to be associated with any known QTL. However RNA data suggested that it is involved in flowering time responses (Figure 2.11). A further investigation of the gene function involved confirmed it contained an F-Box which has been indicated to participate in epistatic regulation (Shahri *et al*., 2014).

Another finding of possible epistatic response can be found when analysing markers from the Mx2 mapping population. A SNP was detected that did not follow any established models of dominance as described in Chapter 3 (Figure 3.18). One of the alleles did appear to increase the variance of flowering for all genotypes and this finding was not detected by conventional analysis tools. Machine learning does not make assumptions on the distribution of alleles; therefore it does not assume any of the dominance models and

this could be the reason why it has the ability to discover new insight without the bias of presumption.

6.1.3 Interpretability

The ability to interpret results from models is a key consideration when selecting a method for building model and data analysis. In order to use the models and analyse the results to support breeding decisions, the results must be interpretable by the breeder. Random forest has the capability to highlight important attributes that have the strongest effect by examining their importance score and to highlight the potential regions in which trait associations were found. Visualisation can provide valuable insight revealing the potential over dominance markers related to yield as illustrated in Chapter 4 (Figure 4.25, Figure 4.26, Figure 4.27, Figure 4.28, and Figure 4.29).

Another potential enhancement of random forest which was not explored in this research is to examine the structures of the trees created. This is difficult due to the large numbers of trees created within each forest. However if a subset of data was selected for investigation or an automated mining process is constructed to identify common patterns among trees. The relationships between certain traits and environmental factors could be further pinpointed and investigated. For example, select a subset of markers to be considered where age is greater than a particular value. The relationship of how genetics and maturity are linked can be revealed. Other studies have suggested similar ideas of extracting more information from random forest models (Touw *et al.,* 2013).

6.2 Major Contributions

In this thesis machine learning has been demonstrated as a method to detect marker trait associations and to associate the relationship between genotype, environment and

phenotype in *Miscanthus*. The developed tool RFQTL was used to analyse the published rice datasets for validation and was able to produce consistent results with conventional QTL analysis methods. The results were also confirmed by aligning markers to homologous QTLs in *Sorghum* genome. Random forest was also used to analyse markers from a GBS study from a selected wild *Miscanthus* genotype to perform marker trait association. The analysis results were confirmed by comparing and aligning with the published *Sorghum* genome. The M5 attribute selection method combined with a linear regression model was developed to study the environmental effects on Miscanthus flowering time.

During the course of this research, several important scientific questions have been raised as presented in the introduction section of this thesis. They will be addressed in the following section.

(1) How computational approaches underpin quantitative genetics research?

Quantitative genetics aims to develop models which link genetic variations with changes in an organism's phenotype. Many approaches exist to develop these types of models, such as QTL mapping and genome wide association studies (GWAS). However statistically based approaches are not able to capture all the variation caused by genetics. Many of these approaches only deal with additive genomic effects and do not consider the variances caused by dominance and interactive effects.

Many machine learning algorithms are non-linear, thus they have the potential to detect smaller effect markers which cannot be discovered using the conventional linear approaches. Machine learning can also mine data through attribute selection. It allows for

only those attributes that affect the response to be used in the model. Several studies have suggested that the use of machine learning in genetic modelling may help improve results (Bernardo *et al.,* 2008; Jannink *et al.,* 2010). A comparison of genomic selection approaches looked at machine learning based methods and showed that they performed as well or better than the statistical methods (Heslot *et al.,* 2012).

In this research the random forest algorithm has been applied to perform QTL mapping and a genome wide study on *Miscanthus.* Those studies have resulted in an array of marker-trait associations. The results were validated against published *Sorghum* traits analysis. Markers with high importance were analysed and mapped to the homologous QTL in *Sorghum* genome.

It has proven, from the results of chapter 3 and chapter 4, that computational approaches such as machine learning can effectively detect regions of the genome that show a quantitative effect on the trait modelled.

(2) Why machine learning can potentially increase the power of prediction on crop modelling to facilitate breeding programmes?

Machine learning has been demonstrated as an effective tool to facilitate crop modelling in this research. In Chapter 5, the effects of environmental factors on flowering time, including drought and increased temperature on *Miscanthus* were modelled and studied.

Through the application of both attribute selection and linear regression, predictive models were developed to predict flowering time for several *Miscanthus* genotypes. The use of cross validation confirmed that the model developed is not overfitted to the dataset on which the model was trained. The resulting model has high correlation between the predicted flowering time and the observed values. The model was then confirmed using a

validation dataset. The correlation between the predicted and observed data is high for both stages of flowering time investigated. For the first flowering stage, flag leaf emergence, the correlation was 0.95 and the second flowering stage, panicle emergence, the correlation was 0.88.

These results have demonstrated the power of machine learning in the development of predictive models. By selecting only those parameters which have significant influence on flowering time, a relatively simple model (linear) can be created to accurately predict flowering time.

In addition, many studies have also made use of machine learning in the prediction of crop performance using genomic data (Heslot *et al.,* 2012) or remote sensing data (Uno *et al.,* 2005; Gutiérrez *et al.,* 2008). Another study used soft computing techniques in order to predict cotton yield (Papageorgiou *et al.,* 2011). To conclude, this research coupled with several published studies have demonstrated the capability of machine learning for crop performance prediction with great potential to assist in the decision making process for a breeding programme.

(3) Would a machine learning/data mining approach be an answer to the association of complex Genotype-by-Phenotype-by-Environment (GxPxE)?

Machine learning was applied to model the dynamic relationship of GxPxE in this research. The first study was illustrated in Chapter 3, where environmental factors were taken into consideration to better account for the impact of environmental variances on flowering time. New regions with significant markers were identified by including environmental variables in the analysis. The year attribute was used to distinguish the environmental variance. After the meteorological data was included in the analysis, the importance score of markers were diminished compared to the analysis when excluding

the environmental factors. This has confirmed that the random forest approach was able to account for the meteorological variances.

Another study was to apply a linear model to predict flowering time based upon water availability and temperature as discussed in Chapter 5. Attribute selection was applied to include the degree day and soil moisture deficit as attributes for the model. A prediction model was created and the results had a good degree of accuracy as illustrated in Chapter 5 (Figure 5.4 and Figure 5.5). This model was further employed to formulate a hypothesis on the effects of drought and changing climatic conditions on flowering time (Figure 5.6, Figure 5.7, and Figure 5.8).

Published studies have also used machine learning to model interactions with environmental effects, such as best watering regimes (Fukuda *et al.,* 2013). Very few studies have looked at using machine learning for associating GxExP relationship. However it has been utilised for modelling of environmental system such as rainfall run off (Dawson & Willby, 1998). These types of models could be improved by increasing the accuracy and resolution of the input data used in GxExP association models.

As demonstrated in the two studies of this research, machine learning/data mining again has proven that it possess ability to detect and learn complex patterns. This makes it a suitable candidate for understanding the complex relationship between genetics and environment.

(4) Will machine learning approach be a better alternative than statistics to dissect the complex traits and conduct high-throughput marker analysis?

Statistics has long been used in dissecting complex traits. Commonly used statistics methods make assumptions that often contradict many established theories of genetics such as the assumptions of only linear (additive genetic) relationships exists (Jannink *et*

al., 2010). Another issue with classical statistical analysis is that it assumes that all attributes are having an even effect on the response.

Phenotypical traits are likely to be controlled by a combination of a few large effect loci combined with many smaller effect loci scattered across the whole genome while some regions will have no effect at all (Buckler *et al*., 2009) The goal of quantitative genetic analysis is to detect all markers related to a trait of interest. Using machine learning through attribute selection can remove those non-effect markers and can fit the model with only those actually having an effect on the trait. This is the origin of the term data mining that machine learning mines data for the informative parameters and discounts all others.

Marker discovery over the past few years have progressed rapidly both in number of markers available and the high throughput methods for genotyping. This results in more processing time needed to handle massive numbers of instances and attributes. Machine learning algorithms are designed to be computationally efficient to handle large quantity of data. Besides, several machine learning algorithms can take advantage of parallelisation to reduce computation time for analysis.

Another advantage of machine learning is its capability of conducting advanced modelling and data representation. For example non-linear models can capture the interactions between genes through better data representation and predictions. Many current methods, such as the linear regression methods (Ogutu *et al.,* 2012*)*, on the other hand, are incapable of detecting interactions between genes and therefore are limited to explain additive effects only. The detection of gene interactions was found during the rice data analysis where new regions were identified using the random forest analysis. Other studies also suggest the advantages of using machine learning in genomic models such

as genomic selection(Bernardo, 2008)

From the conclusion of this research, it is the expected that statistics will still play its role in genetic analysis to dissect the gene but machine learning will provide an enhanced tool for high-throughput marker discovery and dissect complex traits.

(5) Why and how a computational approach can help to drive 21st century breeding programmes?

Modern crop breeding programmes such as those seen in rice, wheat, and maize (Miura *et al.,* 2008; Gupta *et al.,* 2009; Crossa *et al.,* 2013*)* and also in willow (Hanley & Karp, 2013) use a wide range of high throughput genomic resources, novel analysis methodologies and bioinformatics tools to assist in breeding. Marker assisted selection, either through the development of whole genome models (Luan *et al.,* 2009; Sorrells *et al.,* 2011) or through the use of 'significant' markers detected by QTL mapping or GWAS (Prasanna *et a*l., 2010; Steele *et a*l. 2013; Ashraf & Foolad, 2013), has been widely used in modern breeding programmes. Great advances in next generation sequencing has led to the development of economical methods for high density marker generation and high-throughput genotyping (Elshire *et al.,* 2011). The potential to use genetic information to increase the speed of breeding is greater than ever.

Current analytical approaches used are mainly based on linear models such as elastic nets and ridge regression (Ogutu *et al.,* 2012). The computational approach can handle non-linear problems more efficiently.

One such non-linear algorithm is random forest. By creating binary partitions within the problem space the random forest algorithm is capable of modelling both the effects of individual parameters and also the interactions between them. It has been used in this research for QTL detection (Chapter 3) and marker trait associations (Chapter 4) and has

been proven to be a versatile approach for genetic research and breeding.

The ability to accurately predict traits and understand the contributing factors is vitally important in modern breeding programmes. Through the use of the M5 attribute selection and linear regression, a flowering time prediction model was developed to accurately predict flowering time under different future climate scenarios.

Whether it is used for marker trait associations or for crop modelling prediction and knowledge discovery, machine learning is proven to be a powerful tool for modem crop development.

6.3 Future Research

Throughout this research, machine learning has been applied in order to 'mine' a massive amount of both historical and new data collected by the *Miscanthus* breeding and genetic research teams at IBERS. This research has led to the development of several hypotheses on marker trait associations in wild *Miscanthus* germplasm, the detection of QTL controlling flowering time in a *M. sinensis* mapping population, and an understanding of some of the environmental factors which affect flowering time. Furthermore, these studies have started to demonstrate the huge potential of machine learning as a powerful data analysis tool to underpin breeding. The following section will discuss how to apply machine learning further in data analysis and modelling as an intelligent decision making tool for breeding. Some of the unanswered questions raised within these studies and potential future experiments will also be discussed.

6.3.1 Further application of machine learning to underpin breeding

The tools and models developed in this research could easily be adapted and trained

to generate a breeding score for genomic selection (GS). Heslot *et al* (2012) previously suggested the use of random forest as a method for genomic selection, but did recommend caution as it is an untested method for genomic selection.

In order to test the random forest for GS analysis and develop a comprehensive GS model for use in the *Miscanthus* breeding programme, the next focus needs to include progeny in genotyping and data analysis. In order to have an effective GS analysis, progeny of those plants under study should also be phenotyped and genotyped to provide a comprehensive dataset needed to facilitate GS.

Supervised machine learning algorithms require the data sets that include a set of inputs and the responses they produce. In the case of plant breeding, the inputs would be genomic information, such as GBS markers, and the responses would be the phenotypes measured. However thanks to the advance in NGS it is easier to genotype plants than to phenotype them. Therefore, the applications of an intelligent method which can select the minimum requirement of phenotype are desirable.

Active learning is the application of specially designed semi-supervised machine learning algorithms which can interact with a user (Settles, 2010). These algorithms are used when the labelling of data is not economical, usually due to high cost or difficultly in attaining labels with abundant data. In a plant breeding example the labelling of data could refer to either phenotypical or genotypic analysis of a particular genotype.

The active learner can make requests from the user to label particular data points, this could be genotype information about a new group of progeny, or phenotypical data on a particular genotype. The decision of which data point to label is made by some pre-defined heuristic rules. This heuristic will often take into account the cost, either economical or a measure of difficulty in attaining the labels. The learner will attempt to

optimise learning with regards to the cost of requesting labels to minimize cost, but at the same time maintain the greatest knowledge gain.

Active learning can help to improve the efficiency of the GS process. The first stage would be the development of a GS model to predict breeding values. Assuming this has been done and that all or the majority of new crosses can be genotyped, the model will first be used to predict the best progeny to take forward. It will not only provide breeding score predictions but also suggest which genotypes should be used for further evaluation to improve the accuracy of the GS model. The labelling decision will be made based upon an active learning heuristic that takes into account budgetary and manpower constraints, while maximizing the information gain. Phenotype information can then be collected on the selected genotypes and used to improve the GS model.

The differences are subtle between a normal GS model and the one combined with active learning. This difference comes from the active learning based GS model can iteratively retrain based upon the selected data points. Some might argue this is already being done in GS as new progeny are monitored anyway and the models are retrained when accuracy decreases. However computers can create more complex models of cost interactions than a human breeder may be able to.

Earlier we discussed the costs in terms of simply money and time; however this could include the selection of multi-location trial sites. For example if only 60 seeds are available, active learning can help to choose which location should be selected to evaluate or how should they be distributed.

This application of active learning could lead to the development of intelligent breeding systems to provide smart decision support tools for breeders to reduce the cost (or an optimised strategy for an available budget) and also to speed up the development of

new varieties.

For *Mis*canthus, a GS driven programme is still a few years away, due to the requirement to develop more genetic resources, genotyping and phenotyping.

The development of an intelligent breeding assisted programme is one of the main goals for machine learning and will require further research. By integration genomic, phenotypic and environmental data, machine learning based tools can be applied automatically to the newly integrated dataset for further analysis. Active learning based GS can be used to estimate breeding scores and develop new scientific hypothesis that can be further investigated to understand the biological process that control many traits of interest in *Miscanthus.* Advancements in phenomics such as the state-of-art national plant phenomics centre at IBERS will also help to increase the quality of phenotype data.

6.3.2 Genomic Simulations

This research mainly used raw data sets generated from experimental trials. However validating these results empirically is a large scientific undertaking that will either require the development and analysis of many additional populations or through gene discovery and genetic modification to test gene function. This means that many of the markers detected, especially the hundreds of small effect markers are unlikely to be validated empirically.

To test the effectiveness of random forest as a marker analysis method, simulation may provide a solution to resolve this issue. A whole genome simulation model could be developed with a collection of simulated large and small effect QTLs across the entire genome. An array of associated markers and thousands of unrelated markers can be created to test the analysis tool developed. Epistatic relationships could also be built into

this simulated dataset to confirm the ability of nonlinear methods for detecting these relationships. Although a simulation may not be able to accurately reflect the true nature of the relationships, it can verify the ability of the tools developed.

6.3.3 GxE Interaction in *Miscanthus*

Machine learning has been shown to be capable of performing the required analysis to model genotype by environment (GxE) interactions. To develop a more comprehensive model of GxE interactions in *Miscanthus* to make better predictions*,* a multi-location with several replicated trial consisting of a large number of genotypes would be needed to provide more variance. Only the differences between years and not locations were considered in this research. In order to develop models that can account for the variations between locations with different climate conditions, it require the inclusions of additional meteorological parameters into the models developed to improve the prediction quality.

Improved models can be further developed to better understand the effects of many other climate conditions on important phenotypes. Meanwhile, the models can assist in understanding which meteorological factors have the stronger effect on important traits, such as yield, to inform breeding decisions so that various varieties can be developed to suit different climate conditions.

Another important area of research is to fully understand the effects of flowering on yield. Experiments performed in controlled environments are able to discover some of the factors effecting flowering time in *M. sacchariflorus* and demonstrate its effect on yield (Jensen *et al.,* 2013). It is expected that this relationship between genotype and environment can be linked and extrapolated to the field performance through the application of machine learning to pinpoint not only what environmental factors affect

flowering time but also to understand how they affect the resulting yield.

6.3.4 Using computation to understand drought

Drought has been shown to have a negative effect on yield in species such as wheat, rice, and *Miscanthus* (Denčić *et al.,* 2000; Pantuwan *et al.,* 2002; Hasting *et al.,* 2009b). Given that future climate models suggest that droughts are to become more frequent (Olesen & Bindi, 2002; Schröter et al., 2005) therefore there is a need to further study the impact of drought on plant phenotype.

A plants ability to tolerate drought is controlled by many different traits including but not limited to root morphology, leaf rolling and leaf death (Kamoshita *et al.,* 2008). Kamoshita *et al* also demonstrated in this study that morphological differences such as tiller number can change transpiration rates therefore altering the tolerance of a plant. Each of these different traits can have a different magnitude of effect on drought tolerance which is often made up of a combination of several traits. Plants also deal with water stress in a manner which is not easily observed. It requires either specialist tools to measure or must be estimated through modelling. Physiological responses such as stomatal conductance and water use efficiency are examples of the measurements needed (Hufstetler *et al.,* 2007).

Radiation use efficency (RUE) is the ability of a plant to convert radiation into biomass with higher RUE leading to greater biomass accumulation. However under drought RUE is diminished (Yordanov *et al.,* 2000). In *Miscanthus,* RUE has shown to be higher in hybrids (Davey *et al.,* in preparation). Davey *et al* developed a model of RUE for *Miscanthus* however when the model was applied to data collected at the second location, the model over predicted the yield. They suggest that this could be due to a drought effect at the

second location. However attempts by the authors to model the effects of drought on RUE have proven unsuccessful. Computational modelling and automation could be a useful tool that can help to understand the effects of drought on RUE and subsequently achieve higher biomass accumulation.

There are two options to perform crop studies, either in the field or in pot based experiments. In pot based experiments, drought can be controlled whereas in the field experiment drought cannot be directly controlled. However pot experiments often do not accurately reflect the behaviour of the plant in the field. Computational modelling could be a valuable tool to develop association between these two types of experiments.

As mentioned earlier, increasing the efficiency of RUE can potentially increase yield in hybrids. However the impacts of drought on RUE could possibly be great; therefore further study is important. RUE measurements require destructive harvests and demand large numbers of clonal replicates in the trial. In the study of Davey *et al*, RUE was modelled in the field. However, if this was done in a pot experiment, an automated facility could be developed to model the association between RUE and drought without the need for destructive harvest.

By utilising weighing scales to constantly record the weight of plants and an automated watering system, machine learning could be employed to develop a comprehensive model of biomass accumulation under two different conditions, drought and irrigated. Using data from the experiment, the RUE model could be iteratively parameterised. Of course this would require measurements of LAI and light interception that will either have to be performed manually or through the automated systems.

Pot based facilities make it is easier to measure RUE; however as mentioned earlier this may not reflect field conditions. Modelling using machine learning approach could be a

solution to this problem. For example, the artificial neural networks have already been utilised in water run off modelling (Dawson & Wilby, 1998) and could potentially be applied again to develop the spatial model using data collected from reflectometers in the field. The established model could allow for a higher resolution model of soil moisture to be developed and improve the accuracy of the models such as those developed in chapter 5. By increasing the number of genotypes in the field trials can also enhance the association of drought tolerance between genotype and environment. Through a combination of both field and pot experiments, computational modelling could be a powerful tool to help to better understand the biology underlying drought and its effect on the plant performance at a phenotypic and physiological levels.

6.4 Concluding Remarks

Machine learning has been demonstrated in this research as a powerful tool to identify markers and regions of the genome that relate to particular traits of interest. It is still a relatively new approach when compared to statistical analysis. The applications of machine learning are still under exploration by the scientists and its potential application is enormous.

Machine learning, also referred to as data mining, has several advantages over statistics. Through attribute selection it is able to 'mine' data to extract influential attributes, thereby creating simpler models. The concept of learning from data means that the results can be used to better generalise the problem to create hypotheses. These advantages make machine learning more suited to analyse the complex problems and can be a valuable tool for breeders

With more public databases of Omics resources available, the ability to use data from

multiple sources could lead to improvements in breeding. For example, by the application of comparative genomics, breeders can make use of data from other crops to verify the results from their marker-trait associations. The Ondex system is one example of a system which attempts to automatically mine data from multiple sources for data integration (Köhler *et al.,* 2006).

This research has started the process of developing predictive models needed to facilitate the use of genetics research in the *Miscanthus* breeding programme. More studies and further exploration are needed to better establish the associations in progeny and how *Miscanthus* interspecies crossing affects the genome and trait associations.

Machine learning is an effective method for the discovery of marker traits associations and modelling of genotype and environment interaction as evident from this research. It is anticipated the machine learning will play a crucial role in the future of crop breeding. Machine learning also has high potential for intelligent automated data processing to improve the quality and accuracy of data and subsequently enhance the analysis. The ability to perform hypothesis formulation from data is another great strength of machine learning. It provides further scientific insights and advances our understanding of how to bridge the gap between genetics and breeding.

7 References

Adams, K.L. & Wendel, J.F. (2005) Polyploidy and genome evolution in plants. *Current opinion in plant biology*. 8 (2), pp. 135–141.

Agichtein, E., Brill, E. & Dumais, S. (2006) Improving web search ranking by incorporating user behavior information. *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*. pp. 19.

Aha, D., Kibler, D. & Albert, M. (1991) Instance-based learning algorithms. *Machine learning*. 6 (1), pp. 37–66.

Allen, R.G., Pereira, L.S., Raes, D. & Smith, M. (1998) *Crop evapotranspiration: guidelines for computing crop water requirements. Food and Agriculture Organizations of the United Nations, Rome, Italy, 1998*.

OECD-FAO Agricultural (2013) *OECD-FAO Agricultural Outlook 2013*. OECD-FAO Agricultural Outlook. OECD Publishing.

Armstead, I.P., Turner, L.B., Farrell, M., Skøt, L., Gomez, P., Montoya, T., Donnison, I.S., King, I.P. & Humphreys, M.O. (2004) Synteny between a major heading-date QTL in perennial ryegrass (Lolium perenne L.) and the Hd3 heading-date locus in rice. *TAG. Theoretical and applied genetics.* 108 (5), pp. 822–828.

Armstrong, J. (1974) Eclectic Research and Construct Validation. *Models of Buyer Behavior-conceptual,* pp. 3–14.

Ashraf, M. & Foolad, M.R. (2013) Crop breeding for salt tolerance in the era of molecular markers and marker-assisted selection R. Tuberosa (ed.). *Plant Breeding*. 132 (1), pp. 10–20.

Atienza, G., Satovic, Z., Petersen, K., Dolstra, O. & Martín, A. (2002) Preliminary genetic linkage map of Miscanthus sinensis with RAPD markers. *TAG. Theoretical and applied genetics.* 105 (6-7), pp. 946–952.

Atienza, S., Satovic, Z., Petersen, K., Dolstra, O. & Martin, A. (2003a) Identification of QTLs associated with yield and its components in Miscanthus sinensis Anderss. *Euphytica*. 132 (3), pp. 353–361.

Atienza, S.G., Satovic, Z., Petersen, K.K., Dolstra, O. & Martín, a (2003b) Identification of QTLs influencing agronomic traits in Miscanthus sinensis Anderss. I. Total height, flag-leaf height and stem diameter. *TAG. Theoretical and applied genetics.* 107 (1), pp. 123–129.

Atienza, S.G., Satovic, Z., Petersen, K.K., Dolstra, O. & Martin, A. (2003c) Influencing combustion quality in Miscanthus sinensis Anderss.: identification of QTLs for calcium, phosphorus and sulphur content. *Plant Breeding*. 122 (2), pp. 141–145.

Atienza, S.G., Satovic, Z., Petersen, K.K., Dolstra, O. & Martín, A. (2003d) Identification of QTLs influencing combustion quality in Miscanthus sinensis Anderss. II. Chlorine and potassium content. *TAG. Theoretical and applied genetics.* 107 (5), pp. 857–863.

Bai, Y., Luo, L. & Voet, E. (2010) Life cycle assessment of switchgrass-derived ethanol as transport fuel. *The International Journal of Life Cycle Assessment*. 15 (5), pp. 468–477.

Bajpai, P. & Dash, V. (2012) Hybrid renewable energy systems for power generation in stand-alone applications: A review. *Renewable and Sustainable Energy Reviews*. 16 (5), pp. 2926–2939.

Bay, S.D. (2000) Multivariate discretization of continuous variables for set mining.

Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 315–319.

Van Berloo, R. & Stam, P. (1998) Marker-assisted selection in autogamous RIL populations: a simulation study. *TAG Theoretical and Applied Genetics*. 96 (1), pp. 147–154.

Bernardo, R. (2008) Molecular Markers and Selection for Complex Traits in Plants: Learning from the Last 20 Years. *Crop Science*. 48 (5), pp. 1649.

Bernardo, R. (2001) What If We Knew All the Genes for a Quantitative Trait in Hybrid Crops?. *Crop Science*. 41 (1), pp. 1.

Biomass Energy Centre (2013) *UK biomass power stations*. (March), pp. 1–2.

Boichard, D., Guillaume, F., Baur, A., Croiseau, P., Rossignol, M.N., Boscher, M.Y., Druet, T., Genestout, L., Colleau, J.J., Journaux, L., Ducrocq, V. & Fritz, S. (2012) Genomic selection in French dairy cattle. *Animal Production Science*. 52 (3), pp. 115.

Brachi, B., Morris, G.P. & Borevitz, J.O. (2011) Genome-wide association studies in plants: the missing heritability is in the field. *Genome biology*. 12 (10), pp. 232.

Bréda, N.J.J. (2003) Ground-based measurements of leaf area index: a review of methods, instruments and current controversies. *Journal of experimental botany*. 54 (392), pp. 2403–2417.

Breiman, L. (1996) Bagging predictors. *Machine Learning*. 24 (2), pp. 123–140.

Breiman, L. (2001a) Random forests. *Machine learning*. pp. 5–32.

Breiman, L. (2001b) Statistical modeling: The two cultures. *Statistical Science*. 16 (3), pp. 199–231.

Broman, K.W., Wu, H., Sen, S. & Churchill, G. a. (2003) R/qtl: QTL mapping in experimental crosses. *Bioinformatics*. 19 (7), pp. 889–890.

Buckler, E.S., Holland, J.B., Bradbury, P.J., Acharya, C.B., Brown, P.J., Browne, C., Ersoz, E., Flint-Garcia, S., Garcia, A., Glaubitz, J.C., others, Goodman, M.M., Harjes, C., Guill, K., et al. (2009) The genetic architecture of maize flowering time. *Science*. 325 (5941), pp. 714.

Burbidge, R., Rowland, J. & King, R. (2007) Active learning for regression based on query by committee. *Intelligent Data Engineering and Automated Learning - IDEAL 2007*. pp. 209–218.

Cao, A., Xing, L., Wang, X., Yang, X., Wang, W., Sun, Y., Qian, C., Ni, J., Chen, Y., Liu, D., Wang, X. & Chen, P. (2011) Serine/threonine kinase gene Stpk-V, a key member of powdery mildew resistance gene Pm21, confers powdery mildew resistance in wheat. *Proceedings of the National Academy of Sciences of the United States of America*. 108 (19), pp. 7727–7732.

Cerasuolo, M., Richter, G.M., Cunniff, J., Purdy, S., Shield, I. & Karp, A. (2013) A pseudo-3D model to optimise the target traits of light interception in short-rotation coppice willow. *Agricultural and Forest Meteorology*. 173pp. 127–138.

Chang, M., Yih, W. & Meek, C. (2008) Partitioned logistic regression for spam filtering. In: *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. p. pp. 97.

Chen, H., Chung, W., Xu, J.J., Wang, G., Qin, Y. & Chau, M. (2004) Crime data mining: a general framework and some examples. *Computer*. 37 (4), pp. 50–56.

Chen, H., Zeng, D., Atabakhsh, H., Wyzga, W. & Schroeder, J. (2003) COPLINK:

managing law enforcement data and knowledge. *Communications of the ACM*. 46 (1),

Cherubini, F. & Strømman, A.H. (2011) Chemicals from lignocellulosic biomass: opportunities, perspectives, and potential of biorefinery systems. *Biofuels, Bioproducts and Biorefining*. 5 (5), pp. 548–561.

Choi, H.-K., Mun, J.-H., Kim, D.-J., Zhu, H., Baek, J.-M., Mudge, J., Roe, B., Ellis, N., Doyle, J., Kiss, G.B., Young, N.D. & Cook, D.R. (2004) Estimating genome conservation between crop and model legume species. *Proceedings of the National Academy of Sciences of the United States of America*. 101 (43), pp. 15289–15294.

Chou, C.-H. (2009) Miscanthus plants used as an alternative biofuel material: The basic studies on ecology and molecular evolution. *Renewable Energy*. 34 (8), pp. 1908–1912.

Clifton-Brown, J. (2000) Water Use Efficiency and Biomass Partitioning of Three Different Miscanthus Genotypes with Limited and Unlimited Water Supply. *Annals of Botany*. 86 (1), pp. 191–200.

Cobb, J.E., Zaloumis, S.G., Scurrah, K.J., Harrap, S.B. & Ellis, J. a (2010) Evidence for two independent functional variants for androgenetic alopecia around the androgen receptor gene. *Experimental dermatology*. 19 (11), pp. 1026–1028.

Cockram, J., White, J., Zuluaga, D.L., Smith, D., Comadran, J., Macaulay, M., Luo, Z., Kearsey, M.J., Werner, P., Harrap, D., Tapsell, C., Liu, H., Hedley, P.E., Stein, N., et al. (2010) Genome-wide association mapping to candidate polymorphism resolution in the unsequenced barley genome. *Proceedings of the National Academy of Sciences of the United States of America*. 107 (50), pp. 21611–21616.

Combs, E. & Bernardo, R. (2013) Accuracy of Genomewide Selection for Different Traits

with Constant Population Size, Heritability, and Number of Markers. *The Plant Genome*. 6 (1)

Croiseau, P., Legarra, A., Guillaume, F., Fritz, S., Baur, A., Colombani, C., Robert-Granié, C., Boichard, D. & Ducrocq, V. (2011) Fine tuning genomic evaluations in dairy cattle through SNP pre-selection with the Elastic-Net algorithm. *Genetics research*. 93 (6), pp. 409–417.

Cronn, R., Knaus, B.J., Liston, A., Maughan, P.J., Parks, M., Syring, J. V & Udall, J. (2012) Targeted enrichment strategies for next-generation plant biology. *American journal of botany*. 99 (2), pp. 291–311.

Crossa, J., Pérez, P., Hickey, J., Burgueño, J., Ornella, L., Cerón-Rojas, J., Zhang, X., Dreisigacker, S., Babu, R., Li, Y., Bonnett, D. & Mathews, K. (2014) Genomic prediction in CIMMYT maize and wheat breeding programs. *Heredity*. 112 (1), pp. 48–60.

Crowley, T.J. (2000) Causes of Climate Change Over the Past 1000 Years. *Science*. 289 (5477), pp. 270–277.

Cunniff, J. & Cerasuolo, M. (2011) Lighting the way to willow biomass production. *Journal of the science of food and agriculture*. 91 (10), pp. 1733–1736.

Daelemans, W. & Hoste, V. (2002) Evaluation of machine learning methods for natural language processing tasks. In: *LREC 2002 : third international conference on language resources and evaluation*

Darvasi, A., Weinreb, A., Minke, V., Weller, J.I. & Soller, M. (1993) Detecting marker-QTL linkage and estimating QTL gene effect and map location using a saturated genetic map. *Genetics*. 134 (3), pp. 943–951.

Davey, C.L., Jones, L.E., Squance. M., Purdy. S.J., Maddison. A.L., Cunniff. J., Donnison, I., Clifton-Brown, J. (*In preparation*). Radiation capture and conversion efficiencies of *Miscanthus sacchariflorus*, *M. sinensis* and their naturally occurring hybrid *M. x giganteus*.

Dawson, C.W. & Wilby, R. (1998) An artificial neural network approach to rainfall-runoff modelling. *Hydrological Sciences Journal*. 43 (1), pp. 47–66.

Defra (2013) *Area of Crops Grown For Bioenergy in England and the UK : 2008-2011*. (January).

Department of Energy & Climate Change (2011) *UK Renewable Energy Roadmap*. (July).

Denčić, S., Kastori, R., Kobiljski, B. & Duggan, B. (2000) Evaluation of grain yield and its components in wheat cultivars and landraces under near optimal and drought conditions [online]. *Euphytica*. pp. 43–52. [Accessed 12 September 2014].

Deschamps, S., Llaca, V. & May, G.D. (2012) Genotyping-by-Sequencing in Plants. *Biology*. 1 (3), pp. 460–483.

Deuter, M. (2000) Breeding approaches to improvement of yield and quality in Miscanthus grown in Europe. *EMI Project, Final report*. pp. 28–52.

Dodds, D.R. & Gross, R.A. (2007) Chemistry. Chemicals from biomass. *Science (New York, N.Y.)*. 318 (5854), pp. 1250–1251.

Domingos, P. (2012) A few useful things to know about machine learning. *Communications of the ACM*. 55 (10), pp. 78.

Domingos, P. (1999) The Role of Occam's Razor in Knowledge Discovery. *Data Mining and Knowledge Discovery*. 3 (4), pp. 409–425.

Donnison. I., (*In Preparation*) QTL detection in a *Miscanthus sinensis* mapping family

Eller, E., Hawks, J. & Relethford, J.H. (2011) Local Extinction and Recolonization, Species Effective Population Size, and Modern Human Origins. *Human Biology*. 76 (5).

Elshire, R.J., Glaubitz, J.C., Sun, Q., Poland, J. a, Kawamoto, K., Buckler, E.S. & Mitchell, S.E. (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PloS one*. 6 (5), pp. e19379.

Erdinc, O. & Uzunoglu, M. (2012) Optimum design of hybrid renewable energy systems: Overview of different approaches. *Renewable and Sustainable Energy Reviews*. 16 (3), pp. 1412–1425.

ESRI 2011. ArcGIS Desktop: Release 10. Redlands, CA: Environmental Systems Research Institute.

European Parliament and Council of the European Union (2009) *Directive of the European Parliament and of the Council on the Promotion of the Use of Energy from Renewable Source.*

Falconer, D.S. & Mackay, T.F.C. (1996) *Introduction to Quantitative Genetics*.

Farrar, K., Bryant, D.N., Turner, L., Gallagher, J. a., Thomas, A., Farrell, M., Humphreys, M.O. & Donnison, I.S. (2011) Breeding for Bio-ethanol Production in Lolium perenne L.: Association of Allelic Variation with High Water-Soluble Carbohydrate Content. *BioEnergy Research*. 5 (1), pp. 149–157.

Felderhoff, T.J., Murray, S.C., Klein, P.E., Sharma, a., Hamblin, M.T., Kresovich, S., Vermerris, W. & Rooney, W.L. (2012) QTLs for Energy-related Traits in a Sweet × Grain Sorghum. *Crop Science*. 52 (5), pp. 2040

Finckh, M.R., Gacek, E.S., Goyeau, H., Lannou, C., Merz, U., Mundt, C.C., Munk, L., Nadziak, J., Newton, A.C., de Vallavieille-Pope, C. & Wolfe, M.S. (2000) Cereal variety and species mixtures in practice, with emphasis on disease resistance. *Agronomie*. 20 (7), pp. 813–837.

Flavell, R., Cruz, C.H. de B., Christie, M., Allen, J., Keller, M., Gilna, P. & Kell, D.B. (2011) Moving forward with biofuels. *Nature (London)*. 474 (7352), pp. S26–S30.

Welcome Trust Foundation (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*. 447 (7145), pp. 661–678.

Francia, E., Tacconi, G., Crosatti, C., Barabaschi, D., Bulgarelli, D., Dall'Aglio, E. & Valè, G. (2005) Marker assisted selection in crop plants. *Plant Cell, Tissue and Organ Culture*. 82 (3), pp. 317–342.

Frank, E. & Witten, I. (1999) Making better use of global discretization. In: *The Sixteenth International Conference on Machine Learning.*

Franks, S.J., Sim, S. & Weis, A.E. (2007) Rapid evolution of flowering time by an annual plant in response to a climate fluctuation. *Proceedings of the National Academy of Sciences of the United States of America*. 104 (4), pp. 1278–1282.

Fukuda, S., Spreer, W., Yasunaga, E., Yuge, K., Sardsud, V. & Müller, J. (2013) Random Forests modelling for the estimation of mango (Mangifera indica L. cv. Chok Anan) fruit yields under different irrigation regimes. *Agricultural Water Management*. 116pp. 142–150.

Funkhouser, T., Shin, H., Toler-Franklin, C., Castañeda, A.G., Brown, B., Dobkin, D., Rusinkiewicz, S. & Weyrich, T. (2011) Learning how to match fresco fragments.

Journal on Computing and Cultural Heritage. 4 (2), pp. 1–13.

Furbank, R.T. & Tester, M. (2011) Phenomics--technologies to relieve the phenotyping bottleneck. *Trends in plant science*. 16 (12), pp. 635–644.

Gao, J., Suzuki, H. & Yu, B. (2006) Approximation lasso methods for language modeling. In: *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL.* pp. 225–232.

Gauder, M., Graeff-Hönninger, S., Lewandowski, I. & Claupein, W. (2012) Long-term yield and performance of 15 different Miscanthus genotypes in southwest Germany. *Annals of Applied Biology*. 160 (2), pp. 126–136.

Gifford, J.M., Chae, W.B., Swaminathan, K., Moose, S.P. & Juvik, J. a. (2014) Mapping the genome of Miscanthus sinensis for QTL associated with biomass productivity. *GCB Bioenergy.*

GoldenHelix (n.d.) *SNP & Variation Suite*.

Gonza, J., Clifton-brown, J.C., Lewandowski, I., Andersson, B., Basch, G., Christian, D.G., Kjeldsen, J.B., Jørgensen, U., Mortensen, J. V, Riche, A.B., Schwarz, K., Tayebi, K. & Teixeira, F. (2001) Performance of 15 Miscanthus Genotypes at Five Sites in Europe. *Agronomy Journal*. 93 (September-October), pp. 1013–1019.

Gupta, P.K., Langridge, P. & Mir, R.R. (2009) Marker-assisted wheat breeding: present status and future possibilities. *Molecular Breeding*. 26 (2), pp. 145–161.

Gutiérrez, P. a., López-Granados, F., Peña-Barragán, J.M., Jurado-Expósito, M., Gómez-Casero, M.T. & Hervás-Martínez, C. (2008) Mapping sunflower yield as affected by Ridolfia segetum patches and elevation by applying evolutionary product unit neural

networks to remote sensed data. *Computers and Electronics in Agriculture*. 60 (2), pp. 122–132.

Haley, C.S. & Knott, S. a (1992) A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity*. 69 (4), pp. 315–324.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. & Witten, I.H. (2009) The WEKA data mining software. *ACM SIGKDD Explorations Newsletter*. 11 (1), pp. 10.

Hamblin, M.T., Buckler, E.S. & Jannink, J.-L. (2011) Population genetics of genomics-based crop improvement methods. *Trends in genetics : TIG*. 27 (3), pp. 98–106.

Hammer, G.L., Kropff, M.J., Sinclair, T.R. & Porter, J.R. (2002) Future contributions of crop modelling—from heuristics and supporting decision making to understanding genetic regulation and aiding crop improvement. *European Journal of Agronomy*. 18 (1-2), pp. 15–31.

Hanley, S.J. & Karp, A. (2013) Genetic strategies for dissecting complex traits in biomass willows (Salix spp.). *Tree physiology*. pp. 1–14.

Hansen, J., Sato, M., Ruedy, R., Lo, K., Lea, D.W. & Medina-Elizade, M. (2006) Global temperature change. *Proceedings of the National Academy of Sciences of the United States of America*. 103 (39), pp. 14288–14293.

Hart, G.E., Schertz, K.F., Peng, Y. & Syed, N.H. (2001) Genetic mapping of Sorghum bicolor (L.) Moench QTLs that control variation in tillering and other morphological characters. *TAG Theoretical and Applied Genetics*. 103 (8), pp. 1232–1242.

Hastie, T., Tibshirani, R. & Friedman, J. (2009) *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY: Springer New York.

Hastings, A., Clifton-Brown, J., Wattenbach, M., Mitchell, C.P. & Smith, P. (2009a) The development of MISCANFOR, a new Miscanthus crop growth model: towards more robust yield predictions under different climatic and soil conditions. *GCB Bioenergy*. 1 (2), pp. 154–170.

Hastings, A., Clifton-Brown, J., Wattenbach, M., Mitchell, C.P., Stampfl, P. & Smith, P. (2009b) Future energy potential of Miscanthus in Europe. *GCB Bioenergy*. 1 (2), pp. 180–196.

Hayes, B.J., Bowman, P.J., Chamberlain, A.J. & Goddard, M.E. (2009) Invited review: Genomic selection in dairy cattle: Progress and challenges. *Journal of Dairy Science*. 92 (2), pp. 433–443.

Hayes, B.J., Cogan, N.O.I., Pembleton, L.W., Goddard, M.E., Wang, J., Spangenberg, G.C. & Forster, J.W. (2013) Prospects for genomic selection in forage plant species O. A. Rognli (ed.). *Plant Breeding*. 132 (2), pp. 133–143.

Heaton, E.A., Dohleman, F.G., Fernando Miguez, A., Juvik, J.A., Lozovaya, V., Widholm, J., Zabotina, O.A., McIsaac, G.F., David, M.B., Voigt, T.B. & others (2010) Miscanthus: a promising biomass crop. *Advances in Botanical Research*. 56 (10).

Heffner, E.L., Lorenz, A.J., Jannink, J.-L. & Sorrells, M.E. (2010) Plant Breeding with Genomic Selection: Gain per Unit Time and Cost. *Crop Science*. 50 (5), pp. 1681.

Heffner, E.L., Sorrells, M.E. & Jannink, J.-L. (2009) Genomic Selection for Crop Improvement. *Crop Science*. 49 (1), pp. 1.

Heschel, M.S. & Riginos, C. (2005) Mechanisms of selection for drought stress tolerance and avoidance in Impatiens capensis (Balsaminaceae). *American journal of botany*. 92 (1), pp. 37–44.

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

Hill, J., Hambley, M., Forster, T., Mewissen, M., Sloan, T., Scharinger, F., Trew, A. & Ghazal, P. (2008) SPRINT: A new parallel framework for R. *BMC Bioinformatics*. 9 (1), pp. 558.

HIRAYOSHI, I., NISHIKAWA, K. & KATO, R. (1955) Cytogenetical Studies on forage plants. (IV) Self-incompatibility in Miscanthus. *Ikushugaku zasshi*

Hodkinson, T.R., Chase, M.W., Lledó, M.D., Salamin, N. & Renvoize, S.A. (2002a) Phylogenetics of Miscanthus, Saccharum and related genera (Saccharinae, Andropogoneae, Poaceae) based on DNA sequences from ITS nuclear ribosomal DNA and plastid trnLintron and trnL-F intergenic spacers. *Journal of plant research*. 115 (5), pp. 381–392.

Hodkinson, T.R., Chase, M.W., Takahashi, C., Leitch, I.J., Bennett, M.D. & Renvoize, S.A. (2002b) The use of dna sequencing (ITS and trnL-F), AFLP, and fluorescent in situ hybridization to study allopolyploid Miscanthus (Poaceae). *American journal of botany*. 89 (2), pp. 279–286.

Hoerl, A.E. & Kennard, R.W. (1970) Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*. 12 (1), pp. 55–67.

Hoi, S.C.H., Jin, R., Zhu, J. & Lyu, M.R. (2006) Batch mode active learning and its application to medical image classification. *Proceedings of the 23rd international conference on Machine learning - ICML '06*. pp. 417–424.

Huang, X., Zhao, Y., Wei, X., Li, C., Wang, A., Zhao, Q., Li, W., Guo, Y., Deng, L., Zhu, C., Fan, D., Lu, Y., Weng, Q., Liu, K., et al. (2012) Genome-wide association study of

flowering time and grain yield traits in a worldwide collection of rice germplasm. *Nature genetics*. 44 (1), pp. 32–39.

Hufstetler, E.V., Boerma, H.R., Carter, T.E. & Earl, H.J. (2007) Genotypic Variation for Three Physiological Traits Affecting Drought Tolerance in Soybean. *Crop Science*. 47 (1), pp. 25.

Hunter, S., Jones, P., Mitchell, A., Apweiler, R., Attwood, T.K., Bateman, A., Bernard, T., Binns, D., Bork, P., Burge, S., de Castro, E., Coggill, P., Corbett, M., Das, U., et al. (2012) InterPro in 2011: new developments in the family and domain prediction database. *Nucleic acids research*. 40 (Database issue), pp. D306–12.

Hyne, V. & Kearsey, M.J. (1995) QTL analysis: further uses of “marker regression.” *Theoretical and Applied Genetics*. 91 (3), pp. 471–476.

Ings, J., Mur, L. a J., Robson, P.R.H. & Bosch, M. (2013) Physiological and growth responses to water deficit in the bioenergy crop Miscanthus x giganteus. *Frontiers in plant science*. 4 (November), pp. 468.

Izawa, T. (2007) Adaptation of flowering-time by natural and artificial selection in Arabidopsis and rice. *Journal of experimental botany*. 58 (12), pp. 3091–3097.

Jannink, J., Bink, M.C. & Jansen, R.C. (2001) Using complex plant pedigrees to map valuable genes. *Trends in plant science*. 6 (8), pp. 337–342.

Jannink, J.-L., Lorenz, A.J. & Iwata, H. (2010) Genomic selection in plant breeding: from theory to practice. *Briefings in functional genomics*. 9 (2), pp. 166–177.

Jenkins, G., Murphy, J., Sexton, D., Lowe, J., Office, M., & Centre, H. (2010). UK climate projections: Briefing report. Met Office Hadley Centre, Exceter, UK.

Jensen, E., Farrar, K., Thomas-Jones, S., Hastings, A., Donnison, I. & Clifton-Brown, J. (2011a) Characterization of flowering time diversity in Miscanthus species. *GCB Bioenergy*. 3 (5), pp. 387–400.

Jensen, E., Squance, M., Hastings, A., Thomas-Jones, S., Farrar, K., Huang, L., King, R., Clifton-Brown, J., Donnison, I. & Editors (2011b) Understanding the value of hydrothermal time on flowering in Miscanthus species. *Aspects of Applied Biology*. (112), pp. 181–189.

Jensen, E., Robson, P., Norris, J., Cookson, A., Farrar, K., Donnison, I. & Clifton-Brown, J. (2013) Flowering induction in the bioenergy grass Miscanthus sacchariflorus is a quantitative short-day response, whilst delayed flowering under long days increases biomass accumulation. *Journal of experimental botany*. 64 (2), pp. 541 552.

Jin, J. & Jiang, C. (2002) Spatial variability of soil nutrients and site-specific nutrient management in the P.R. China. *Computers and Electronics in Agriculture*. 36 (2-3), pp. 165–172.

Jørgensen, U. (2011) Benefits versus risks of growing biofuel crops: the case of Miscanthus. *Current Opinion in Environmental Sustainability*. 3 (1-2), pp. 24–30.

Kaastra, I. & Boyd, M. (1996) Designing a neural network for forecasting financial and economic time series. *Neurocomputing*. 10 (3), pp. 215–236.

Kamoshita, A., Babu, R.C., Boopathi, N.M. & Fukai, S. (2008) Phenotypic and genotypic analysis of drought-resistance traits for development of rice cultivars adapted to rainfed environments. *Field Crops Research*. 109 (1-3), pp. 1–23.

Kang, H.M., Sul, J.H., Service, S.K., Zaitlen, N.A., Kong, S.-Y., Freimer, N.B., Sabatti, C. & Eskin, E. (2010) Variance component model to account for sample structure in

genome-wide association studies. *Nature genetics*. 42 (4), pp. 348–354.

Kaplan, S. a (2003) Boosted decision tree analysis of surface-enhanced laser desorption/ionization mass spectral serum profiles discriminates prostate cancer from noncancer patients. *The Journal of urology*. 169 (4), pp. 1620.

Karl, T.R. & Trenberth, K.E. (2003) Modern global climate change. *Science (New York, N.Y.)*. 302 (5651), pp. 1719–1723.

Karp, A., Hanley, S.J., Trybush, S.O., Macalpine, W., Pei, M. & Shield, I. (2011) Genetic improvement of willow for bioenergy and biofuels. *Journal of integrative plant biology*. 53 (2), pp. 151–165.

Kawahara, Y., de la Bastide, M., Hamilton, J.P., Kanamori, H., McCombie, W.R., Ouyang, S., Schwartz, D.C., Tanaka, T., Wu, J., Zhou, S., Childs, K.L., Davidson, R.M., Lin, H., Quesada-Ocampo, L., et al. (2013) Improvement of the Oryza sativa Nipponbare reference genome using next generation sequence and optical map data. *Rice*. 6 (1), pp. 4.

Kearsey, M. (1998) The principles of QTL analysis (a minimal mathematics approach). *Journal of Experimental Botany*. 49 (327), pp. 1619–1623.

Kearsey, M.J. & Farquhar, A.G.L. (1998) *Short Review QTL analysis in plants ; where are we now ?* 80 (October 1997), pp. 137–142.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R. & Salzberg, S.L. (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome biology*. 14 (4), pp. R36.

Kingsolver, J.G., Hoekstra, H.E., Hoekstra, J.M., Berrigan, D., Vignieri, S.N., Hill, C.E.,

Hoang, A., Gibert, P. & Beerli, P. (2001) The strength of phenotypic selection in natural populations. *The American naturalist*. 157 (3), pp. 245–261.

Knight, C.G., Platt, M., Rowe, W., Wedge, D.C., Khan, F., Day, P.J.R., McShea, A., Knowles, J. & Kell, D.B. (2009) Array-based evolution of DNA aptamers allows modelling of an explicit sequence-fitness landscape. *Nucleic acids research*. 37 (1), pp. e6.

Knott, S. a, Elsen, J.M. & Haley, C.S. (1996) Methods for multiple-marker mapping of quantitative trait loci in half-sib populations. *TAG. Theoretical and applied genetics.* 93 (1-2), pp. 71–80.

Köhler, J., Baumbach, J., Taubert, J., Specht, M., Skusa, A., Rüegg, A., Rawlings, C., Verrier, P. & Philippi, S. (2006) Graph-based analysis and visualization of experimental results with ONDEX. *Bioinformatics (Oxford, England)*. 22 (11), pp. 1383–1390.

Kononenko, I. & Hong, S.J. (1997) Attribute selection for modelling [online]. *Future Generation Computer Systems*. 13 (2-3), pp. 181–195.

Kumar, N., Kulwal, P.L., Balyan, H.S. & Gupta, P.K. (2006) QTL mapping for yield and yield contributing traits in two mapping populations of bread wheat. *Molecular Breeding*. 19 (2), pp. 163–177.

Kumar, S. & Subramanian, S. (2002) Mutation rates in mammalian genomes. *Proceedings of the National Academy of Sciences of the United States of America*. 99 (2), pp. 803–808.

Kump, K.L., Bradbury, P.J., Wisser, R.J., Buckler, E.S., Belcher, A.R., Oropeza-Rosas, M. a, Zwonitzer, J.C., Kresovich, S., McMullen, M.D., Ware, D., Balint-Kurti, P.J. &

Holland, J.B. (2011) Genome-wide association study of quantitative resistance to southern leaf blight in the maize nested association mapping population. *Nature genetics*. 43 (2), pp. 163–168.

Larranaga, P. (2006) Machine learning in bioinformatics. *Briefings in Bioinformatics*. 7 (1), pp. 86–112.

Laurie, C.C., Chasalow, S.D., LeDeaux, J.R., McCarroll, R., Bush, D., Hauge, B., Lai, C., Clark, D., Rocheford, T.R. & Dudley, J.W. (2004) The genetic architecture of response to long-term artificial selection for oil concentration in the maize kernel. *Genetics*. 168 (4), pp. 2141–2155.

LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W. & Jackel, L.D. (1989) Backpropagation applied to handwritten zip code recognition. *Neural computation*. 1 (4), pp. 541–551.

Lemus, R. & Lal, R. (2005) Bioenergy Crops and Carbon Sequestration. *Critical Reviews in Plant Sciences*. 24 (1), pp. 1–21.

Lewandowski, I. & Clifton-Brown, J. (2000) Miscanthus: European experience with a novel energy crop. *Biomass and Bioenergy* 19 (2000), pp. 209–227.

Lewandowski, I. & Heinz, A. (2003) Delayed harvest of miscanthus—influences on biomass quantity and quality and environmental impacts of energy production. *European Journal of Agronomy*. 19 (1), pp. 45–63.

Li, H., Peng, Z., Yang, X., Wang, W., Fu, J., Wang, J., Han, Y., Chai, Y., Guo, T., Yang, N., Liu, J., Warburton, M.L., Cheng, Y., Hao, X., et al. (2013) Genome-wide association study dissects the genetic architecture of oil biosynthesis in maize kernels. *Nature genetics*. 45 (1), pp. 43–50.

Li, Z.K., Luo, L.J., Mei, H.W., Wang, D.L., Shu, Q.Y., Tabien, R., Zhong, D.B., Ying, C.S., Stansel, J.W., Khush, G.S. & Paterson, a H. (2001) Overdominant epistatic loci are the primary genetic basis of inbreeding depression and heterosis in rice. I. Biomass and grain yield. *Genetics*. 158 (4), pp. 1737–1753.

Liaw, A. & Wiener, M. (2002) Classification and Regression by randomForest. *R News*. 2 (3), pp. 18–22.

Limure, T., Kihara, M., Ichikawa, S., Ito, K., Takeda, K. & Sato, K. (2011) Development of DNA markers associated with beer foam stability for barley breeding. *Theoretical and applied genetics.* 122 (1), pp. 199–210.

Lin, Y., Schertz, K. & Paterson, A. (1995) Comparative Analysis of QTLs Affecting Plant Height and Maturity Across the Poaceae, in Reference to an Interspecific Sorghum Population. *Genetics*.

Lipka, A.E., Tian, F., Wang, Q., Peiffer, J., Li, M., Bradbury, P.J., Gore, M. a, Buckler, E.S. & Zhang, Z. (2012) GAPIT: genome association and prediction integrated tool. *Bioinformatics (Oxford, England)*. 28 (18), pp. 2397–2399.

Liu, H., Bayer, M., Druka, A., Russell, J.R., Hackett, C. a, Poland, J., Ramsay, L., Hedley, P.E. & Waugh, R. (2014) An evaluation of genotyping by sequencing (GBS) to map the Breviaristatum-e (ari-e) locus in cultivated barley. *BMC genomics*. 15 (1), pp. 104.

Lobell, D.B. & Burke, M.B. (2010) On the use of statistical models to predict crop yield responses to climate change. *Agricultural and Forest Meteorology*. 150 (11), pp. 1443–1452.

Luan, T., Woolliams, J.A., Lien, S., Kent, M., Svendsen, M. & Meuwissen, T.H.E. (2009) The accuracy of genomic selection in Norwegian red cattle assessed by cross-

validation. *Genetics*. 183 (3), pp. 1119–1126.

Lukens, L. & Doebley, J. (1999) Epistatic and environmental interactions for quantitative trait loci involved in maize evolution. *Genetical Research*. pp. 291–302.

Ma, L., Wiggans, G.R., Wang, S., Sonstegard, T.S., Yang, J., Crooker, B. a, Cole, J.B., Van Tassell, C.P., Lawlor, T.J. & Da, Y. (2012a) Effect of sample stratification on dairy GWAS results. *BMC genomics*. 13pp. 536.

Ma, X.-F., Jensen, E., Alexandrov, N., Troukhan, M., Zhang, L., Thomas-Jones, S., Farrar, K., Clifton-Brown, J., Donnison, I., Swaller, T. & Flavell, R. (2012b) High resolution genetic mapping by genome sequencing reveals genome duplication and tetraploid genetic structure of the diploid Miscanthus sinensis. *PloS one*. 7 (3), pp. e33821.

Mackay, I. & Powell, W. (2007) Methods for linkage disequilibrium mapping in crops. *Trends in plant science*. 12 (2), pp. 57–63.

Madhumati, B. (2014) Potential and application of molecular markers techniques for plant genome analysis. *Int. J. Pure App. Biosci*. 2 (1), pp. 169–188.

Magerman, D.M. (1995) Statistical decision-tree models for parsing. In: *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*. pp. pp. 276–283.

Manolio, T. a, Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L. a, Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., Cho, J.H., Guttmacher, A.E., Kong, A., Kruglyak, L., et al. (2009) Finding the missing heritability of complex diseases. *Nature*. 461 (7265), pp. 747–753.

Marcelis, L.F., Heuvelink, E. & Goudriaan, J. (1998) Modelling biomass production and yield of horticultural crops: a review. *Scientia Horticulturae*. 74 (1-2), pp. 83–111.

Marx, V. (2013) Biology: The big challenges of big data. *Nature*. 498 (7453), pp. 255–260.

McVicker, I.F.G. (1946) The calculation and use of degree-days. *The Building services engineer*.

Meehl, G.A., Zwiers, F., Evans, J., Knutson, T., Mearns, L. & Whetton, P. (2000) Trends in Extreme Weather and Climate Events: Issues Related to Modeling Extremes in Projections of Future Climate Change. *Bulletin of the American Meteorological Society*. 81 (3), pp. 427–436.

Meinke, D.W., Cherry, J.M., Dean, C., Rounsley, S.D. & Koornneef, M. (1998) Arabidopsis thaliana: a model plant for genome analysis. *Science (New York, N.Y.)*. 282 (5389), pp. 662, 679–682.

Metzker, M.L. (2010) Sequencing technologies - the next generation. *Nature reviews. Genetics*. 11 (1), pp. 31–46.

Meuwissen, T.H., Hayes, B.J. & Goddard, M.E. (2001) Prediction of total genetic value using genome-wide dense marker maps. *Genetics*. 157 (4), pp. 1819–1829.

Miedaner, T. & Korzun, V. (2012) Marker-assisted selection for disease resistance in wheat and barley breeding. *Phytopathology*. 102 (6), pp. 560–566.

Miguez, F.E., Maughan, M., Bollero, G. a. & Long, S.P. (2012) Modeling spatial and dynamic variation in growth, yield, and yield stability of the bioenergy crops Miscanthus × giganteus and Panicum virgatum across the conterminous United States. *GCB Bioenergy*. 4 (5), pp. 509–520.

Miura, K., Ashikari, M. & Matsuoka, M. (2011) The role of QTLs in the breeding of high-yielding rice. *Trends in plant science*. 16 (6), pp. 319–326.

Mitchell, T.M. (1997). *Machine learning*. McGraw-Hill series in computer science. McGraw-Hill.

Mohan, M., Nair, S., Bhagwat, A. & Krishna, T. (1997) Genome mapping, molecular markers and marker-assisted selection in crop plants. *Molecular Breeding*. pp. 87–103.

Montes, J.M., Melchinger, A.E. & Reif, J.C. (2007) Novel throughput phenotyping platforms in plant genetic studies. *Trends in plant science*. 12 (10), pp. 433–436.

Morris, G.P., Ramu, P., Deshpande, S.P., Hash, C.T., Shah, T., Upadhyaya, H.D., Riera-Lizarazu, O., Brown, P.J., Acharya, C.B., Mitchell, S.E., Harriman, J., Glaubitz, J.C., Buckler, E.S. & Kresovich, S. (2013) Population genomic and genome-wide association studies of agroclimatic traits in sorghum. *Proceedings of the National Academy of Sciences of the United States of America*. 110 (2), pp. 453–458.

Narum, S.R., Buerkle, C.A., Davey, J.W., Miller, M.R. & Hohenlohe, P. a (2013) Genotyping-by-sequencing in ecological and conservation genomics. *Molecular ecology*. 22 (11), pp. 2841–2847.

Neumann, K., Kobiljski, B., Denčić, S., Varshney, R.K. & Börner, A. (2010) Genome-wide association mapping: a case study in bread wheat (Triticum aestivum L.). *Molecular Breeding*. 27 (1), pp. 37–58.

Ngai, E.W.T., Hu, Y., Wong, Y.H., Chen, Y. & Sun, X. (2011) The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision Support Systems*. 50 (3), pp. 559–569.

Ogutu, J.O., Schulz-Streeck, T. & Piepho, H.-P. (2012) Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their

extensions. *BMC proceedings*. 6 Suppl 2 (Suppl 2), pp. S10.

Olesen, J.E. & Bindi, M. (2002) Consequences of climate change for European agricultural productivity, land use and policy. *European Journal of Agronomy*. 16 (4), pp. 239–262.

Van Ooijen, J.W. (2004) MapQTL 5, Software for the mapping of quantitative trait loci in experimental populations.

Van Ooijen, J.W. (2011). Multipoint maximum likelihood mapping in a full-sib family of an outbreeding species. *Genetics Research, accepted: June 2011,* published: (2011) 93, 5, 343-349.

Ornella, L. & Tapia, E. (2010) Supervised machine learning and heterotic classification of maize (Zea mays L.) using molecular marker data. *Computers and Electronics in Agriculture*. 74 (2), pp. 250–257.

Pacala, S. & Socolow, R. (2004) Stabilization wedges: solving the climate problem for the next 50 years with current technologies. *Science (New York, N.Y.)*. 305 (5686), pp. 968–972.

Pantuwan, G., Fukai, S., Cooper, M., Rajatasereekul, S. & O'Toole, J.. (2002) Yield response of rice (Oryza sativa L.) genotypes to different types of drought under rainfed lowlands. *Field Crops Research*. 73 (2-3), pp. 153–168.

Papageorgiou, E.I., Markinos, a. T. & Gemtos, T. a. (2011) Fuzzy cognitive map based approach for predicting yield in cotton crop production as a basis for decision support system in precision agriculture application. *Applied Soft Computing*. 11 (4), pp. 3643–3657.

Pasam, R.K., Sharma, R., Malosetti, M., van Eeuwijk, F. a, Haseneyer, G., Kilian, B. &

Graner, A. (2012) Genome-wide association studies for agronomical traits in a world wide spring barley collection. *BMC plant biology*. 12 (1), pp. 16.

Paterson, a H., Schertz, K.F., Lin, Y.R., Liu, S.C. & Chang, Y.L. (1995) The weediness of wild plants: molecular analysis of genes influencing dispersal and persistence of johnsongrass, Sorghum halepense (L.) Pers. *Proceedings of the National Academy of Sciences of the United States of America*. 92 (13), pp. 6127–6131.

Paterson, A.H., Bowers, J.E., Bruggmann, R., Dubchak, I., Grimwood, J., Gundlach, H., Haberer, G., Hellsten, U., Mitros, T., Poliakov, A., Schmutz, J., Spannagl, M., Tang, H., Wang, X., et al. (2009) The Sorghum bicolor genome and the diversification of grasses. *Nature*. 457 (7229), pp. 551–556.

Perlich, C., Dalessandro, B., Raeder, T., Stitelman, O. & Provost, F. (2013) Machine learning for targeted display advertising: transfer learning in action. *Machine Learning*. 95 (1), pp. 103–127.

Pimentel, D., Marklein, A., Toth, M. a., Karpoff, M., Paul, G.S., McCormack, R., Kyriazis, J. & Krueger, T. (2008) Biofuel Impacts on World Food Supply: Use of Fossil Fuel, Land and Water Resources. *Energies*. 1 (2), pp. 41–78.

Pogson, M. (2011) Modelling Miscanthus yields with low resolution input data. *Ecological Modelling*. 222 (23-24), pp. 3849–3853.

Poland, J. a. & Rife, T.W. (2012) Genotyping-by-Sequencing for Plant Breeding and Genetics. *The Plant Genome Journal*. 5 (3), pp. 92.

Prasanna, B.M., Pixley, K., Warburton, M.L. & Xie, C.-X. (2010) Molecular marker-assisted breeding options for maize improvement in Asia. *Molecular Breeding*. 26 (2), pp. 339–356.

Punta, M., Coggill, P.C., Eberhardt, R.Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J., Heger, A., Holm, L., Sonnhammer, E.L.L., Eddy, S.R., et al. (2012) The Pfam protein families database. *Nucleic acids research*. 40 (Database issue), pp. D290–301.

Quinlan, J. (1986) Induction of decision trees. *Machine learning*. pp. 81–106.

R Core Team (2013) *R: A Language and Environment for Statistical Computing*. Available from: http://www.r-project.org.

Rafalski, A. (2002) Applications of single nucleotide polymorphisms in crop genetics. *Current opinion in plant biology*. 5 (2), pp. 94–100.

Recknagel, F., French, M., Harkonen, P. & Yabunaka, K.-I. (1997) Artificial neural network approach for modelling and prediction of algal blooms. *Ecological Modelling*. 96 (1-3), pp. 11–28.

Ritter, K.B., Jordan, D.R., Chapman, S.C., Godwin, I.D., Mace, E.S. & Lynne McIntyre, C. (2008) Identification of QTL for sugar-related traits in a sweet × grain sorghum (Sorghum bicolor L. Moench) recombinant inbred population. *Molecular Breeding*. 22 (3), pp. 367–384.

Robson, P., Jensen, E., Hawkins, S., White, S.R., Kenobi, K., Clifton-Brown, J., Donnison, I. & Farrar, K. (2013) Accelerating the domestication of a bioenergy crop: identifying and modelling morphological targets for sustainable yield increase in Miscanthus. *Journal of Experimental Botany*.

Rosenzweig, C., Iglesias, A., Yang, X., Epstein, P. & Chivian, E. (2001) Climate change and extreme weather events Implications for food production, plant diseases, and pests. *Global change & human health*. 2 (2), pp. 90–104.

Rossum, G. van (1995) *Python tutorial, Technical Report CS-R9526*.

Rosyara, U.R., Gonzalez-Hernandez, J.L., Glover, K.D., Gedye, K.R. & Stein, J.M. (2009) Family-based mapping of quantitative trait loci in plant breeding populations with resistance to Fusarium head blight in wheat as an illustration. *TAG. Theoretical and applied genetics.* 118 (8), pp. 1617–1631.

Sakai, H., Lee, S.S., Tanaka, T., Numa, H., Kim, J., Kawahara, Y., Wakimoto, H., Yang, C., Iwamoto, M., Abe, T., Yamada, Y., Muto, A., Inokuchi, H., Ikemura, T., et al. (2013) Rice Annotation Project Database (RAP-DB): an integrative and interactive database for rice genomics. *Plant & cell physiology*. 54 (2), pp. e6.

Sakai, H., Mizuno, H., Kawahara, Y., Wakimoto, H., Ikawa, H., Kawahigashi, H., Kanamori, H., Matsumoto, T., Itoh, T. & Gaut, B.S. (2011) Retrogenes in rice (Oryza sativa L. ssp. japonica) exhibit correlated expression with their source genes. *Genome biology and evolution*. 3pp. 1357–1368.

Saruta, K., Hirai, Y., Tanaka, K., Inoue, E., Okayasu, T. & Mitsuoka, M. (2013) Predictive models for yield and protein content of brown rice using support vector machine. *Computers and Electronics in Agriculture*. 99pp. 93–100.

Sathya, R. & Abraham, A. (2013) Comparison of Supervised and Unsupervised Learning Algorithms for Pattern Classification. *International Journal of Advanced Research in Artificial Intelligence*. 2 (2), pp. 34–38.

Schein, A.I. & Ungar, L.H. (2007) Active learning for logistic regression: an evaluation. *Machine Learning*. 68 (3), pp. 235–265.

Schlimmer, J.C. & Granger, R.H. (1986) Incremental learning from noisy data. *Machine Learning*. 1 (3), pp. 317–354.

Schröter, D., Cramer, W., Leemans, R., Prentice, I.C., Araújo, M.B., Arnell, N.W., Bondeau, A., Bugmann, H., Carter, T.R., Gracia, C. a, de la Vega-Leinert, A.C., Erhard, M., Ewert, F., Glendining, M., et al. (2005) Ecosystem service supply and vulnerability to global change in Europe. *Science (New York, N.Y.)*. 310 (5752), pp. 1333–1337.

Scott, I.M., Vermeer, C.P., Liakata, M., Corol, D.I., Ward, J.L., Lin, W., Johnson, H.E., Whitehead, L., Kular, B., Baker, J.M., others, Walsh, S., Dave, A., Larson, T.R., et al. (2010) Enhancement of Plant Metabolite Fingerprinting by Machine Learning. *Plant physiology*. 153 (4), pp. 1506.

Scott, L.J., Mohlke, K.L., Bonnycastle, L.L., Willer, C.J., Li, Y., Duren, W.L., Erdos, M.R., Stringham, H.M., Chines, P.S., Jackson, A.U., Prokunina-Olsson, L., Ding, C.-J., Swift, A.J., Narisu, N., et al. (2007) A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science (New York, N.Y.)*. 316 (5829), pp. 1341–1345.

Semel, Y., Nissenbaum, J., Menda, N., Zinder, M., Krieger, U., Issman, N., Pleban, T., Lippman, Z., Gur, A. & Zamir, D. (2006) Overdominant quantitative trait loci for yield and fitness in tomato. *Proceedings of the National Academy of Sciences of the United States of America*. 103 (35), pp. 12981–12986.

Settles, B. (2010) Active learning literature survey. *University of Wisconsin, Madison*.

Shafiee, S., Minaei, S., Moghaddam-Charkari, N. & Barzegar, M. (2014) Honey characterization using computer vision system and artificial neural networks. *Food chemistry*. 159pp. 143–150.

Shahri, W. & Tahir, I. (2014) Flower senescence: some molecular aspects. *Planta*. 239 (2), pp. 277–297.

Sharma, N., Sharma, P., Irwin, D. & Shenoy, P. (2011) Predicting solar generation from weather forecasts using machine learning. *2011 IEEE International Conference on Smart Grid Communications (SmartGridComm)*. pp. 528–533.

Shekoofa, A., Emam, Y., Shekoufa, N., Ebrahimi, M. & Ebrahimie, E. (2014) Determining the most important physiological and agronomic traits contributing to maize grain yield through machine learning algorithms: a new avenue in intelligent agriculture. *PloS one*. 9 (5), pp. e97288.

Sherrington, C. & Moran, D. (2010) Modelling farmer uptake of perennial energy crops in the UK. *Energy Policy*. 38 (7), pp. 3567–3578.

Shinozaki, K., Yamaguchi-Shinozaki, K. & Seki, M. (2003) Regulatory network of gene expression in the drought and cold stress responses. *Current Opinion in Plant Biology*. 6 (5), pp. 410–417.

Shiringani, A.L., Frisch, M. & Friedt, W. (2010) Genetic mapping of QTLs for sugar-related traits in a RIL population of Sorghum bicolor L. Moench. *TAG. Theoretical and applied genetics.* 121 (2), pp. 323–336.

Skøt, L., Humphreys, M.O., Armstead, I., Heywood, S., Skøt, K.P., Sanderson, R., Thomas, I.D., Chorlton, K.H. & Hamilton, N.R.S. (2005) An association mapping approach to identify flowering time genes in natural populations of Lolium perenne (L.). *Molecular Breeding*. 15 (3), pp. 233–245.

Slavov, G., Robson, P., Jensen, E., Hodgson, E., Farrar, K., Allison, G., Hawkins, S., Thomas-Jones, S., Ma, X.-F., Huang, L., Swaller, T., Flavell, R., Clifton-Brown, J. & Donnison, I. (2013) Contrasting geographic patterns of genetic variation for molecular markers vs. phenotypic traits in the energy grass Miscanthus sinensis. *GCB*

Bioenergy. 5 (5), pp. 562–571.

Slavov, G.T., Nipper, R., Robson, P., Farrar, K., Allison, G.G., Bosch, M., Clifton-Brown, J.C., Donnison, I.S. & Jensen, E. (2014) Genome-wide association studies and prediction of 17 traits related to phenology, biomass and cell wall composition in the energy grass Miscanthus sinensis. *The New phytologist*. 201 (4), pp. 1227–1239.

Smith, a. B., Cullis, B.R. & Thompson, R. (2005) The analysis of crop cultivar breeding and evaluation trials: an overview of current mixed model approaches. *The Journal of Agricultural Science*. 143 (06), pp. 449.

Snape, J.W., Law, C.N. & Worland, a J. (1977) Whole chromosome analysis of height in wheat. *Heredity*. 38 (1), pp. 25–36.

Somerville, C., Youngs, H., Taylor, C., Davis, S.C. & Long, S.P. (2010) Feedstocks for lignocellulosic biofuels. *Science (New York, N.Y.)*. 329 (5993), pp. 790–792.

Sorrells, M.E., Heffner, E.L. & Jannink, J.L. (2011) Genomic Selection Accuracy using Multifamily Prediction Models in a Wheat Breeding Program. *The Plant Genome*. 4 (1), pp. 65–75.

Steele, K. a, Price, a H., Witcombe, J.R., Shrestha, R., Singh, B.N., Gibbons, J.M. & Virk, D.S. (2013) QTLs associated with root traits increase yield in upland rice when transferred through marker-assisted selection. *TAG. Theoretical and applied genetics.* 126 (1), pp. 101–108.

Sultan, S.E. (2000) Phenotypic plasticity for plant development, function and life history. *Trends in plant science*. 5 (12), pp. 537–542.

Summerfield, R.J., Roberts, E.H., Ellis, R.H. & Lawn, R.J. (1991) Towards the reliable

prediction of time to flowering in six annual crops. I. The development of simple models for fluctuating field environments. *Experimental Agriculture*. 27 (01), pp. 11–31.

Swaminathan, K., Alabady, M.S., Varala, K., De Paoli, E., Ho, I., Rokhsar, D.S., Arumuganathan, A.K., Ming, R., Green, P.J., Meyers, B.C., Moose, S.P. & Hudson, M.E. (2010) Genomic and small RNA sequencing of Miscanthus x giganteus shows the utility of sorghum as a reference genome sequence for Andropogoneae grasses. *Genome biology*. 11 (2), pp. R12.

Tanner, L., Schreiber, M., Low, J.G.H., Ong, A., Tolfvenstam, T., Lai, Y.L., Ng, L.C., Leo, Y.S., Thi Puong, L., Vasudevan, S.G., Simmons, C.P., Hibberd, M.L. & Ooi, E.E. (2008) Decision tree algorithms predict the diagnosis and outcome of dengue fever in the early phase of illness. *PLoS neglected tropical diseases*. 2 (3), pp. e196.

Taylor, J., King, R.D., Altmann, T. & Fiehn, O. (2002) Application of metabolomics to plant genotype discrimination using statistics and machine learning. *Bioinformatics*. 18 (Suppl 2), pp. S241–S248.

Thuillet, A.-C., Bataillon, T., Poirier, S., Santoni, S. & David, J.L. (2005) Estimation of long-term effective population sizes through the history of durum wheat using microsatellite data. *Genetics*. 169 (3), pp. 1589–1599.

Tian, F., Bradbury, P.J., Brown, P.J., Hung, H., Sun, Q., Flint-Garcia, S., Rocheford, T.R., McMullen, M.D., Holland, J.B. & Buckler, E.S. (2011) Genome-wide association study of leaf architecture in the maize nested association mapping population. *Nature genetics*. 43 (2), pp. 159–162.

Tong, S. & Koller, D. (2002) Support vector machine active learning with applications to

text classification. *The Journal of Machine Learning Research*. 2pp. 45–66.

Tooker, J.F. & Frank, S.D. (2012) Genotypically diverse cultivar mixtures for insect pest management and increased crop yields. *Journal of Applied Ecology*. 49 (5), pp. 974–985.

Touw, W.G., Bayjanov, J.R., Overmars, L., Backus, L., Boekhorst, J., Wels, M. & van Hijum, S. a F.T. (2013) Data mining in the Life Sciences with Random Forest: a walk in the park or lost in the jungle?. *Briefings in bioinformatics*. 14 (3), pp. 315–326.

Trangmar, B.B., Yost, R.S., Wade, M.K., Uehara, G. & Sudjadi, M. (1987) Spatial Variation of Soil Properties and Rice Yield on Recently Cleared Land. *Soil Science Society of America Journal*. 51 (3), pp. 668.

Trapnell, C., Williams, B. a, Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J. & Pachter, L. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology*. 28 (5), pp. 511–515.

Turner, S., Armstrong, L.L., Bradford, Y., Carlson, C.S., Crawford, D.C., Crenshaw, A.T., de Andrade, M., Doheny, K.F., Haines, J.L., Hayes, G., Jarvik, G., Jiang, L., Kullo, I.J., Li, R., et al. (2011) Quality control procedures for genome-wide association studies. *Current protocols in human genetics / editorial board, Jonathan L. Haines.*

Ungerer, M.C., Halldorsdottir, S.S., Purugganan, M.D. & Mackay, T.F.C. (2003) Genotype-environment interactions at quantitative trait loci affecting inflorescence development in Arabidopsis thaliana. *Genetics*. 165 (1), pp. 353–365.

Uno, Y., Prasher, S.O., Lacroix, R., Goel, P.K., Karimi, Y., Viau, a. & Patel, R.M. (2005) Artificial neural networks to predict corn yield from Compact Airborne Spectrographic

Imager data. *Computers and Electronics in Agriculture*. 47 (2), pp. 149–161.

Valentine, J., Clifton-Brown, J., Hastings, A., Robson, P., Allison, G. & Smith, P. (2012) Food vs. fuel: the use of land for lignocellulosic “next generation” energy crops that minimize competition with primary food production. *GCB Bioenergy*. 4 (1), pp. 1–19.

Vales, M.I., Schön, C.C., Capettini, F., Chen, X.M., Corey, a E., Mather, D.E., Mundt, C.C., Richardson, K.L., Sandoval-Islas, J.S., Utz, H.F. & Hayes, P.M. (2005) Effect of population size on the estimation of QTL: a test using resistance to barley stripe rust. *TAG. Theoretical and applied genetics.* 111 (7), pp. 1260–1270.

Vermerris, W. (2008) *Genetic Improvement of Bioenergy Crops*. Wilfred Vermerris (ed.). New York, NY: Springer New York.

Visscher, P.M., Brown, M. a, McCarthy, M.I. & Yang, J. (2012) Five years of GWAS discovery. *American journal of human genetics*. 90 (1), pp. 7–24.

De Vries, S.C., van de Ven, G.W.J., van Ittersum, M.K. & Giller, K.E. (2010) Resource use efficiency and environmental performance of nine major biofuel crops, processed by first-generation conversion techniques. *Biomass and Bioenergy*. 34 (5), pp. 588–601.

Wang, D., Zhu, J., Li, Z. & Paterson, A. (1999) Mapping QTLs with epistatic effects and QTL× environment interactions by mixed linear model approaches. *Theoretical and Applied Genetics*. pp. 1255–1264.

Wang S., C. J. Basten, and Z.-B. Zeng (2012). Windows QTL Cartographer 2.5. Department of Statistics, North Carolina State University, Raleigh, NC. (http://statgen.ncsu.edu/qtlcart/WQTLCart.htm)

Warmuth, M.K., Liao, J., Rätsch, G., Mathieson, M., Putta, S. & Lemmen, C. (2003) Active

learning with support vector machines in the drug discovery process. *Journal of chemical information and computer sciences*. 43 (2), pp. 667–673.

Warrick, a. W. & Gardner, W.R. (1983) Crop yield as affected by spatial variations of soil and irrigation. *Water Resources Research*. 19 (1), pp. 181–186.

Watson, D.J. (1947) Comparative physiological studies on the growth of field crops. *Annals of Botany*. 11 (1), pp. 41–76.

Whittaker, J., Thompson, R. & Denham, M. (2000) Marker-assisted selection using ridge regression. *Genetic Resources*. (75), pp. 249–252.

Widrow, B. & Hoff, M. (1960) Adaptive switching circuits. *IRE WESCON Convention Record*. 4pp. 96–104.

Witten, I.H. & Frank, E. (2000) *Data mining: practical machine learning tools and techniques with Java implementations*. The Morgan Kaufmann series in data management systems. Morgan Kaufmann.

Wu, T.T., Chen, Y.F., Hastie, T., Sobel, E. & Lange, K. (2009) Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics (Oxford, England)*. 25 (6), pp. 714–721.

Xing, Y. & Zhang, Q. (2010) Genetic and molecular bases of rice yield. *Annual review of plant biology*. 61pp. 421–442.

Yano, M., Katayose, Y., Ashikari, M., Yamanouchi, U., Monna, L., Fuse, T., Baba, T., Yamamoto, K., Umehara, Y., Nagamura, Y. & Sasaki, T. (2000) Hd1, a major photoperiod sensitivity quantitative trait locus in rice, is closely related to the Arabidopsis flowering time gene CONSTANS. *The Plant cell*. 12 (12), pp. 2473–2484.

Yonemaru, J., Yamamoto, T., Fukuoka, S., Uga, Y., Hori, K. & Yano, M. (2010) Q-TARO: QTL Annotation Rice Online Database. *Rice*. 3 (2-3), pp. 194–203.

Yordanov, I., Velikova, V. & Tsonev, T. (2000) Plant Responses to Drought, acclimation and stress tolerance. *Photosynthetica*.

Yu, J., Pressoir, G., Briggs, W.H., Vroh Bi, I., Yamasaki, M., Doebley, J.F., McMullen, M.D., Gaut, B.S., Nielsen, D.M., Holland, J.B., Kresovich, S. & Buckler, E.S. (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature genetics*. 38 (2), pp. 203–208.

Zhang, D., Guo, H., Kim, C., Lee, T.-H., Li, J., Robertson, J., Wang, X., Wang, Z. & Paterson, A.H. (2013) CSGRqtl, a comparative quantitative trait locus database for Saccharinae grasses. *Plant physiology*. 161 (2), pp. 594–599.

Zhao, K., Tung, C.-W., Eizenga, G.C., Wright, M.H., Ali, M.L., Price, A.H., Norton, G.J., Islam, M.R., Reynolds, A., Mezey, J., McClung, A.M., Bustamante, C.D. & McCouch, S.R. (2011) Genome-wide association mapping reveals a rich genetic architecture of complex traits in Oryza sativa. *Nature communications*. 2pp. 467.

Zhu, J. (2004) Classification of gene microarrays by penalized logistic regression. *Biostatistics*. 5 (3), pp. 427–443.

Zou, G., Zhai, G., Feng, Q., Yan, S., Wang, A., Zhao, Q., Shao, J., Zhang, Z., Zou, J., Han, B. & Tao, Y. (2012) Identification of QTLs for eight agronomically important traits using an ultra-high-density map based on SNPs generated from high-throughput sequencing in sorghum under contrasting photoperiods. *Journal of experimental botany*. 63 (15), pp. 5451–5462.

8 Appendix

MiscanPheno – Android App for Phenotyping Data Collection

One of the major difficulties of phenotyping is the recording of data in the field, commonly used options are pen and paper or basic mobile computers such as 'Huskys' or simple hand held computers, often the type which run version of the Windows mobile platform with Excel used for data recording. These interfaces are not user friendly when on a small screen, with stylus's often being used to move between cells, or the need to remember to skip a field or move on after each one.

Recent developments in mobile computing, in both phones and tablets, have opened up a whole range of affordable and powerful hand held devices that are now ubiquitous through out the modern world. These new devices are much more capable than their ancestors with better quality screens and internet connectivity. With these advantages it was decided to build an Android App to facilitate the collection of data in the field.

Developed using the Android Java API the first version of the app consisted of a simple front end which allowed a plant id to be first entered, then a series of phenotypes were entered. Upon completion observations were recorded into a flat file system on the device. This version also have capabilities to load field plans, which where downloaded as a csv from the breeding programme database. Field plans were then visualised. A list of plants to be phenotyped could also be loaded and these would then be highlighted on the field plan visualisation. Data could be retrieved from the app using an export function which created a csv file which contained the data collected.

Once a field plan has been loaded it can be 'walked' where each plant is given the

option of being phenotyped or skipped, meaning that each plant's id does not need to be entered by hand if the user is following the field layout (Figure 8.1).

The second version of the app was a simple modification to the phenotypes that could be recorded, and a secondary app was created from the base of the original to perform flowering observations as used in the mapping family 2013 observations to increase speed of collection.

Currently in its third version the app has undergone many changes to make it not only more useful for the data collection in this project, but also a powerful tool for the whole breeding program.

The old flat file system was replaced, the app now uses SQLite to store data (Figure 8.2.) The app talks to the central group database to retrieve field plans instead of having to be download as csv and transferred to the app via JSON. Several direct data feeds now come from the database including a list of phenotypes that can be collected. The use of an SQL database opens up options such as revisiting observations taken, for example if an error was made in data entry. If phenotyping needs to take place over several days, due to large trials and numbers of observations, a margin can be set that means yesterdays phenotypes are still visible if the observation store has not been cleared. Phenotype profiles can now been created in which a user selects the subset phenotypes they wish to record. Once a profile has been selected only those fields in the profile are presented to the user to enter data (Figure 8.3).

A new feature is the ability to record plants as dead, which generates a list in a pre formatted file that can then be used to update the central database. Data is still exported to csv. The csv format was used as it would allow for data checking before uploading to the database, whereas a direct link might mean that errors are not discovered. The csv output

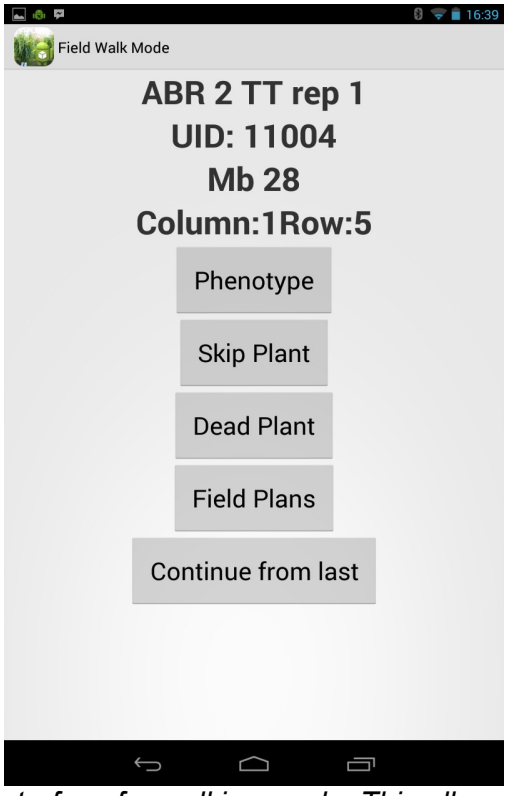

Figure 8.1: Interface for walking mode. This allows a user to chose what they wish to do for each plant in the field, phenotype, skip or record it as dead. There is also an option to jump to the last plant or view the field plan

is now formatted so that it can be automatically loaded into the database, therefore reducing the effort needed to put data into the central database.

Field plan visualisation can still be performed, but thanks to the increased power from using a SQL database dead and phenotyped plants can be visualised in real time from the internal database (Figure 8.4).

An assumption was made that internet may not be available at all trials sites. Therefore field plan data needs to be preloaded before heading to the field, but this data is cached in the devices SQL database. There is also functionality to clear the internal database, the recommendation is that this should be performed between each different trial, after the data has been exported to reduce load on the tablet, and save battery life.

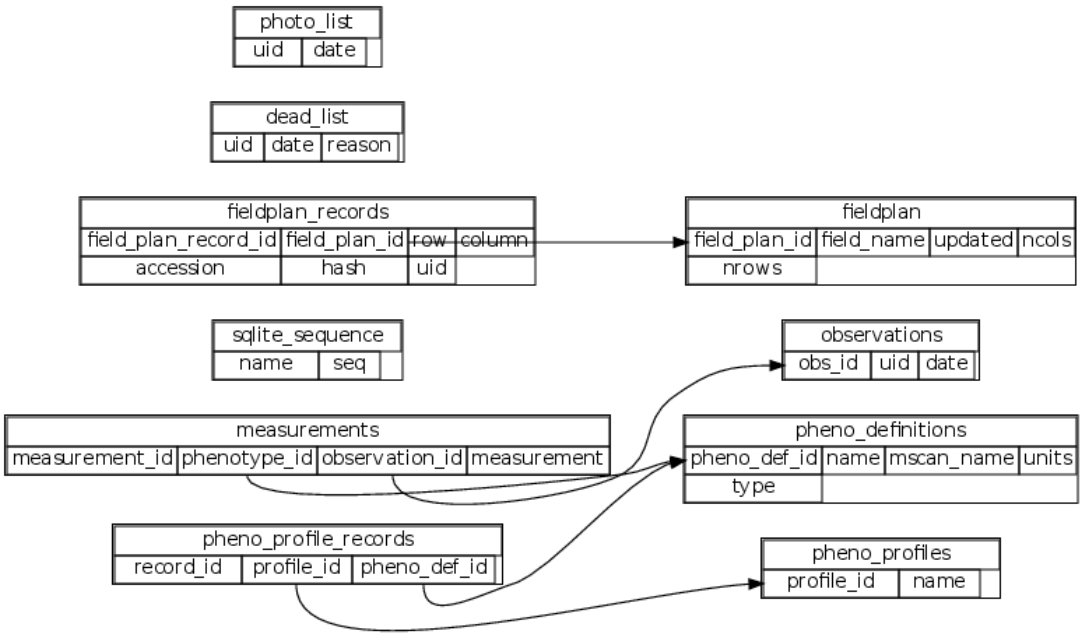

Figure 8.2: Database structure from the MiscanPheno app

One advantage of using a modern technology is that new devices can be used such as bluetooth barcode scanners. Each plant in the *Miscanthus* breeding program has a unique id and this is printed onto a plant label along with a 2D barcode containing the same information. By using a scanner the risk of user error when entering ids disappears, which can be helpful when processing batches of samples, such as fresh and dry weights from subsamples taken at harvest.

Some features that are planned or are under development include a bluetooth receipt printer for recording a hard copy of phenotypes for destructive measurements that cannot be repeated. Also a link up to a weighting scale (most likely using bluetooth) so manual data entry of weights need not be performed, reducing the potential for user error and also leading to an increase in data collection rate, especially when performing batch weighing, such as dry weights out of the oven.

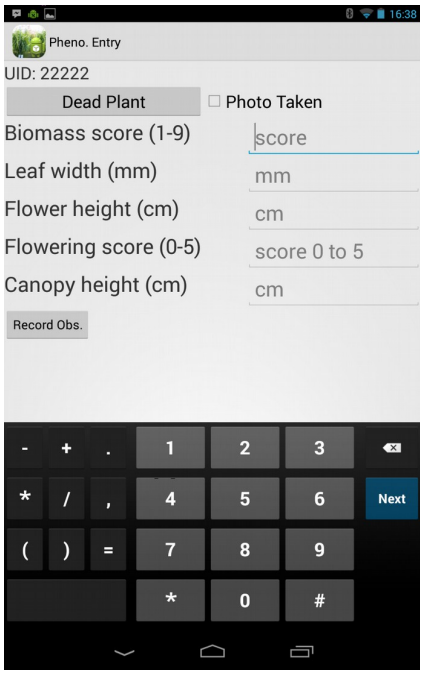

Figure 8.3: Shown is the data entry interface. A user can enter any or all of the fields. Data is automatically recorded into the database after entry.

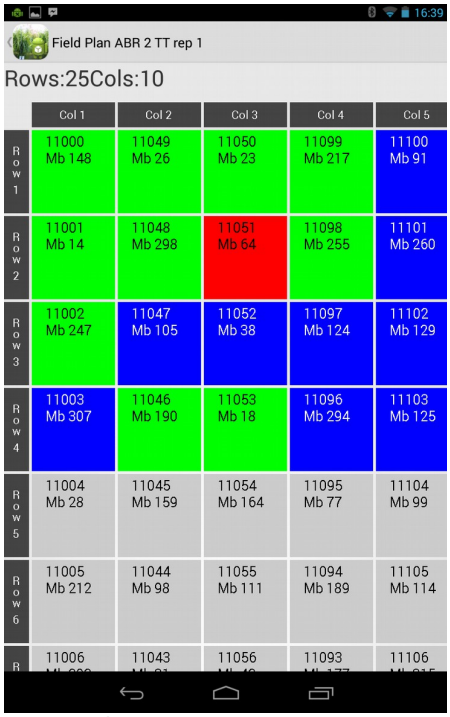

Figure 8.4: Shown is the visualisation of the field plan, it is colour coded to show the current phenotyping state of the field. In this version red is a dead plant, green is phenotyped, and blue is skipped. Colours for each state are user definable.

Database Structure

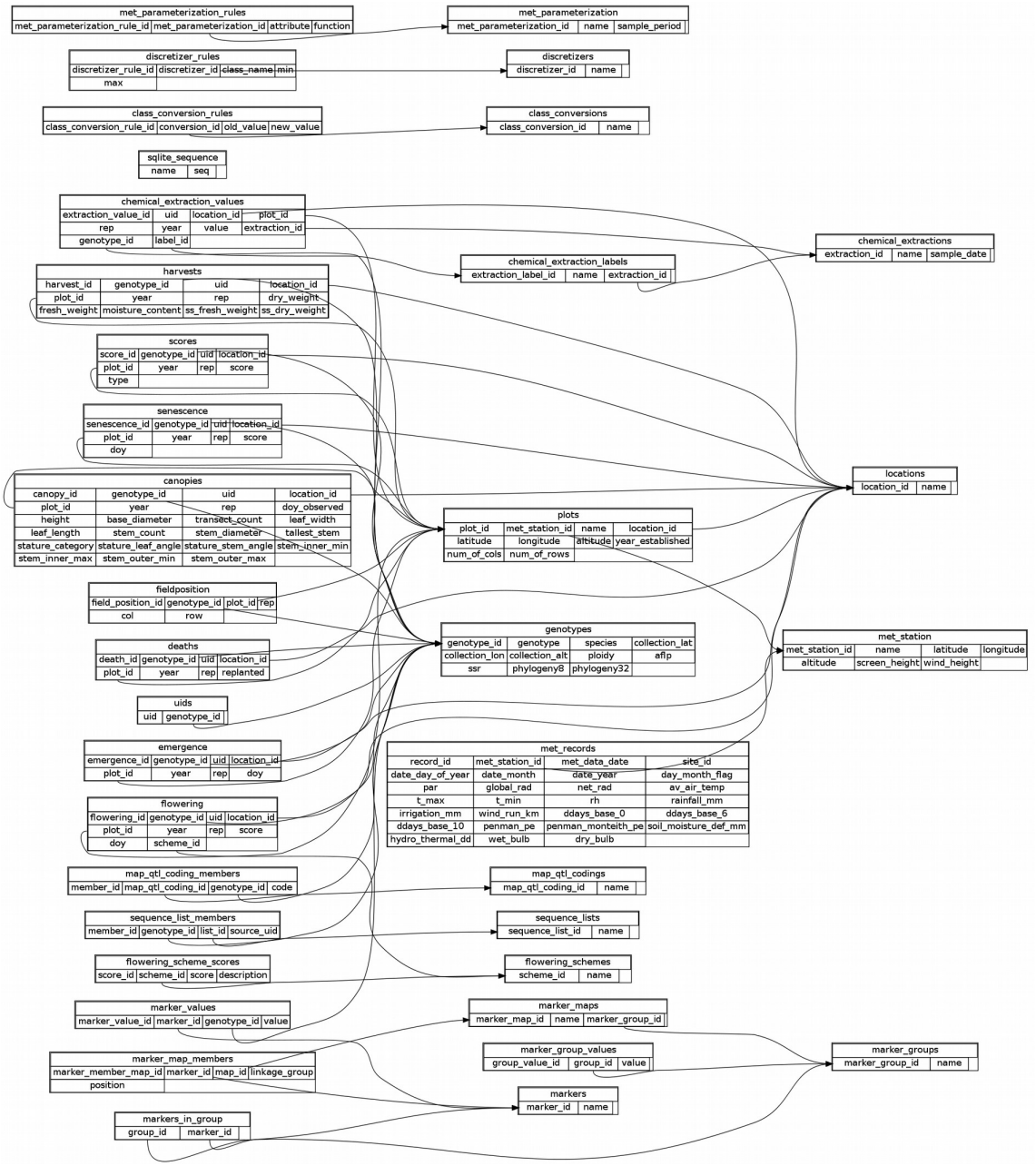

Figure 8.5: Structure of the python interfaced SQLite database. This database was used to hold the genetic, environmental and phenotypical data used throughout this thesis.