GENOMIC SELECTION IN MISCANTHUS SINENSIS

PREDICTION OF BIOMASS QUALITY TRAITS FOR BIOBASED END USES

Colbers, B.M.A.

MSc. Thesis

GENOMIC SELECTION IN MISCANTHUS SINENSIS

PREDICTION OF BIOMASS QUALITY TRAITS FOR BIOBASED END USES

MSc Thesis Plant Breeding
PBR-80436 | 36 ECTS
Submission: March 13, 2018

Colbers, B.M.A. (Bram)
950721-160-090
Master Plant Biotechnology (MPB)

Supervision and Examination:
Andres Torres and Luisa Trindade

Plant Breeding, Wageningen University and Research
Research group: Biobased Economy

September 2017 - March 2018

ABSTRACT

The perennial grass species *Miscanthus sinensis* has been identified as a promising lignocellulosic biomass feedstock that can facilitate the transition towards a biobased economy. Given that *M. sinensis* is still an orphan crop, considerable efforts will be needed to increase the speed and effectiveness of *M. sinensis* breeding. The aims of this study were to characterize the phenotypic variation within the WUR miscanthus collection, obtain insights into the genetic relationships amongst its accessions and to assess the potential of genomic selection to accelerate *M. sinensis* breeding efforts. The accessions within the collection displayed a high degree of phenotypic variation with high estimates of heritability for all 14 traits that were measured during its 5th and 6th growing season. Although sequencing data of 94 *M. sinensis* genotypes had a lower quality than expected, genotypes could be well distinguished based on their genetic relationships. Genomic prediction based on 2600 SNPs resulted in an average prediction accuracy of 0.51. Changing the input of the model and its parameters lead to trait-dependent changes in prediction accuracy. Therefore, optimal settings and recommendations for future experimental design were given. Considering the long establishment phase of *M. sinensis,* it is expected that implementation of genomic selection will substantially increase the rate of genomic improvement for *M. sinensis*. Although sequencing quality was suboptimal, the findings and pipeline generated in this project will guide future research when high quality sequence data becomes available.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 BACKGROUND

1.1 BIOENERGY FROM LIGNOCELLULOSICS

In April 2016, representatives of 174 countries signed the Paris Agreement, a global agreement on the reduction of climate change. In this agreement, countries committed to restrict the increase of global temperature to well below 2 °C above pre-industrial averages, with an aim for a 1.5 °C increase (UNFCCC, 2015). In order to meet the ambitious reductions of climate change and greenhouse gases, novel sustainable sources of energy will be needed to replace fossil energy and reduce carbon emissions. Energy derived from biomass has been identified as a sustainable, renewable form of energy that will be important for improving the sustainability of our modern production, transportation and consumption levels (Fuss et al., 2014). These improvements should lead to a transition from our current fossil-resource intensive economy towards a biobased economy. This concept, the biobased economy, comprises all technological developments allowing for a significant replacement of fossil energy carriers by biomass (Trindade et al., 2010).

Lignocellulosic biomass can be used to produce several forms of energy, for example for heat, liquid fuels, electricity and chemicals. As for liquid fuels, the current practice is the growth of lignocellulosic biomass and using the cellulose and hemicellulose for the production of bioethanol (Trindade et al., 2010). Cellulose and hemicellulose are two major polymers in the plant cell wall, and their fraction of the biomass is the major component of dried biomass. The conversion success of biomass to biofuel is dependent on the saccharification efficiency of converting these polysaccharides into fermentable monosaccharides. Other cell wall components such as lignin can reduce the accessibility of these polymers and thus reduce the saccharification efficiency (Allison et al., 2010; Lewandowski et al., 2016; Zhao et al., 2012).

Current production of biofuels using the cellulose and hemicellulose of lignocellulosic crops is considered a step in the right direction (Stöcker, 2008). However, to create a biobased economy which is commercially successful, all plant biomass needs to be entirely valorised. This would require so called 'integrated biorefineries', where the plant biomass is refined into a broad range of products that cover the operational costs of deconstructing biomass and its transformation into added-value products. Analogous to the petrochemical industry, an integrated biorefinery might produce a high volume, low value liquid transportation fuel and multiple low volume, high value biochemicals that enhance the profitability of the biorefinery (Figure 1). Conceptually, it generates its own energy, which reduces the costs and improves the sustainability of the integrated biorefinery (Speight, 2011).

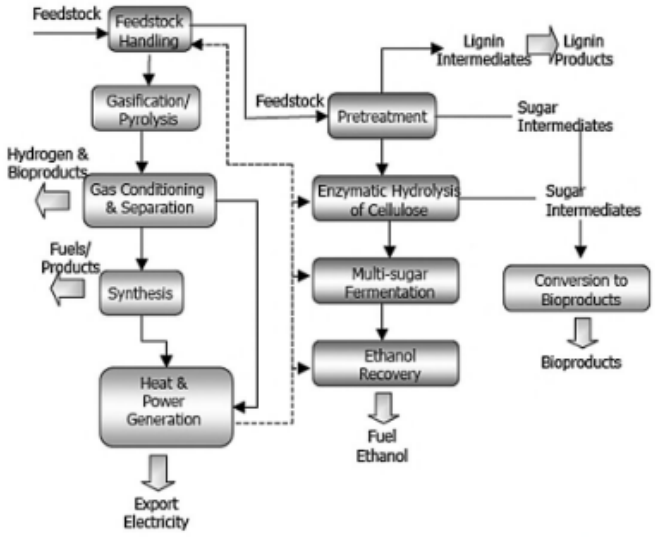


FIGURE 1: A CONCEPT BIOREFINERY FOR LIGNOCELLULOSIC FEEDSTOCKS, ADAPTED FROM SPEIGHT, J.G. (2011)

Following this transition towards the employment of integrated biorefineries, conventional concepts about processing techniques and waste streams need to be revised. An iconic example of an outdated concept is the utilization of lignin. Conventionally, lignin was considered to be a waste stream with a major impact on the recalcitrance of biomass (i.e. resistance of cell walls to enzymatic deconstruction). Novel conversion pathways have shown possibilities to start the deconstruction of the cell wall with the harvest of lignin from intact biomass and subsequently utilize lignin for the production of high value biochemicals. This has given a new meaning to the concept of biomass recalcitrance, which now has been defined as the features of biomass that increase the costs, complexity and energy requirements of operations in the biorefinery and reduce the conversion efficiency of biomass into the desired products (McCann and Carpita, 2015). Under this new definition, biomass recalcitrance is a trait that is end-goal dependent. At a more general level, the new definition implies that the specific features of lignocellulosic biomass composition are not fixed and should be tailored to the demands of specific industrial end-uses.

1.2 MISCANTHUS (SINENSIS)

Although lignocellulosic biomass is considered a promising candidate for providing a sustainable, high yielding feedstock, relatively limited time and effort has been put into breeding and improving lignocellulosic plants (Allwright and Taylor, 2016). The goal of achieving high yields of lignocellulosic biomass is to maximize the amount of biobased products that can be obtained from a given amount of land. In the near future, most of the soils will be needed for food production. Therefore, agricultural soils used for biobased purposes must be highly productive. In the previous section biomass recalcitrance and biomass composition have been addressed as important features of lignocellulosic biomass crops. Other plant characteristics that have been identified as advantageous for biomass crops are a high photosynthetic efficiency, increased recycling of nutrients at end of growing season, high water use efficiency and a high stress tolerance (Taylor et al., 2016).
A category of plants that complies with several of these advantageous characteristics mentioned above are C4 plants. Compared to the more common C3 plants, C4 plants use an improved photosynthetic system that prevents photorespiration. The improved photosynthetic system of C4 plants increases both their energy efficiency and water use efficiency, which results in relative higher yields compared to C3 plants (Ghannoum et al., 2011). Examples of these plants are maize, sugar cane and sorghum.

Several C4 plants that are good candidate species for the production of lignocellulosic biomass belong to the plant genus *Miscanthus*, here further referred to as miscanthus. Miscanthus is a perennial grass native to East Asia, which has adapted to different environments over a wide climatic range and is therefore well suited for the production of biomass under European climatic conditions (Lewandowski et al., 2016). Miscanthus has a high biomass yield, little input demands and a high photosynthetic efficiency. It has an early spring emergence, a long vegetative phase and its production period is 10 -25 years. Miscanthus recycles its nutrients between growing seasons by transporting them to its rhizomes after flowering, which makes the plant very resource-use efficient. (Lewandowski et al., 2016; van der Weijde et al., 2017c). Furthermore, because miscanthus grows equally well on marginal lands, it does not compete with food security (Allwright and Taylor, 2016). Plant developmental processes such as flowering and senescence could also be interesting targets for breeding, as they are important in promoting nutrient remobilization and therefore improve biomass quality (Clifton-Brown and Lewandowski, 2002; Jensen et al., 2017). However, transition from the vegetative phase to the reproductive phase diverts the energy for growing biomass to flowering (Jensen et al., 2011). A combination of a late but promptly and synchronous senescence, that leaves enough time to remobilize plant nutrients to the rhizome, would be ideal.

Within Europe, Miscanthus research has mostly focussed on three miscanthus species, *M. sacchariflorus (2n=4×=76; 4.4 Gb), M. sinensis (2n=2×=38; 5.4 Gb)* and *M. x giganteus (2n=3x=57; 6.8 Gb)*, the latter one being a natural cross between the former two. *M. x giganteus* is the only genotype that is currently commercially available for biomass production. Approximately 20.000 ha is commercially grown in Europe (Lewandowski et al., 2016). In order to improve the interest in, and increase the production area of miscanthus, several issues need to be overcome. Firstly, the triploid *M. x giganteus* is sterile and its propagation has to be performed via rhizomes or *in vitro*. This makes the initial establishment of *M. x giganteus* very expensive. Additionally, the current application of Miscanthus biomass is usage as solid fuel (e.g. pellets) for heat and power generation, which is a low value use. Breeding for traits that valorise miscanthus biomass (e.g. high saccharification efficiency) in *M. x giganteus* is significantly limited because of its sterile nature (van der Weijde et al., 2017b). This drawback is even more apparent in the context of integrated biorefineries and breeding for traits tailored to end-goal dependent qualities. This severely limits the potential of *M. x giganteus* for application in different biobased industries. For this reason, plant breeders have taken interest in the fertile species *M. sacchariflorus* and *M. sinensis*.

1.3 PHENOTYPIC VARIANCE

In order for *M. sacchariflorus* and *M. sinensis* to be used in breeding programs, an important requisite would be to have ample phenotypic variation for important agronomical traits, such as yield and biomass quality. Lewandoski et al. (2016) discovered that *M. sacchariflorus*, *M. sinensis* and novel hybrids from the former species all had genotypes that were able to outperform the commercially grown *M. x giganteus*, which has an average yield of 22 ton DM ha^{-1} yr^{-1} (Heaton et al., 2004). Other research showed that both *M. sinensis* and *M. sacchariflorus* had genotypes with large genotypic variation in cell wall composition. Hence, they can be used in breeding programs to improve biomass quality (Allison et al., 2011; Zhao et al., 2014). Breeding efforts for better cell wall quality and saccharification efficiency have already shown a potential to significantly improve the value of miscanthus biomass (Lewandowski et al., 2016). In comparison to *M. sacchariflorus*, *Miscanthus sinensis* can grow in a larger geographical range and possesses a larger variance in traits related to cold tolerance and cell wall quality (Hodkinson et al., 2015; van der Weijde et al., 2017c). Additionally, the diploid *M. sinensis* facilitates genetic research relative to the tetraploid *M. sacchariflorus*, since practices such as allele fixing and recombination studies will be less complex (Hodkinson et al., 2015). For these reasons, breeding for *M. sinensis* has gained a higher priority compared to *M. sacchariflorus.*

From biochemical analysis on *M. sinensis* was reported that approximately 85% of dry biomass consisted of cell wall material, of which 46% was cellulose and 31% were hemicellulosic polysaccharides (Lewandowski et al., 2016). Other reports show cell wall components range from 28–49% for cellulose, 24–32% for hemicellulose and 15–28% for lignin content (Hodgson et al., 2010; Zhao et al., 2012). Field trials consisting of 8 diverse *M. sinensis* genotypes showed biogas yields from 441 to 520 ml/g dry matter. However, genotypes with a high conversion efficiency were not the most high yielding genotypes. It was concluded that breeding in the future should focus on combining both the high-yielding and high biomass quality traits (Lewandowski et al., 2016). In a F2 mapping population from parents with contrasting cell wall composition it was shown that biomass quality traits had a high heritability. Most biomass quality traits had a H^2 higher than 0.5, and the highest H^2 of 0.62-0.72 was observed for lignin content in the cell wall (ADL/cw) (van der Weijde et al., 2017b). In other words, the potential to move forward in this direction is feasible.

Recent characterization of the Wageningen University & Research (WUR) miscanthus collection showed a wide range of phenotypical variation for various morphological and biochemical traits (Bogers, 2017). The coefficient of variation (CV) for morphological characteristics ranged from 10.4% for 'date of flowering initiation' to 120.7% for stem angle, with an average CV of 36.7%. The CV for cell wall traits ranged from 5.8% for cellulose to 46.8% for 'glucose + xylose yield', with an average CV of 14.9%. Ample variation was found for lignin content with a CV of 12.8%.

Analysis of trait combinations showed examples of genotypes with preferable trait combinations, indicating that breeding for certain trait combinations is possible. For example, plants were found that had a combination of a high yield and a small length, a combination of a high stem yield and either an early or a late flowering, and a combination of a high cellulose conversion and either a low or a high lignin content. Hence, it can be concluded that the Wageningen UR miscanthus collection has a potential for breeding for biobased end uses.

1.4 EXISTING MOLECULAR TOOLS FOR M. SINENSIS

Although *M. sinensis* has preferable characteristics and a high variance in traits important in the context of integrated biorefineries, it has unique genetics and breeding objectives that hamper the usage of classical breeding strategies. *M. sinensis* is self-incompatible and therefore has a highly heterozygous genome. The genome is large, complex and has undergone genome wide duplication (Ma et al., 2012). Furthermore, *M. sinensis* has a long establishment period and it can take two or three years before biomass quality and biomass production are predictive of their values at full maturity (Arnoult et al., 2015; van der Weijde et al., 2017a). The combination of these unique qualities addresses the necessity for molecular tools in Miscanthus breeding.

In the past decades, various techniques for dissecting the genotypic determinants of phenotypic traits have been implemented in breeding efforts. These include the construction of genetic maps, quantitative trait locus (QTL) analyses and genome wide association studies (GWASs). Up until now, several (incomplete) genetic maps have been published. In 2002, Atienza et al. used RAPD markers to identify 28 linkage groups. This map was incomplete since *M. sinensis* has 19 chromosomes. Nevertheless, the map has been useful for the discovery of a number of QTLs (Atienza et al., 2002). In 2012, Swaminathan et al. used 868 RNAseq-markers to produce a framework genetic map with 19 linkage groups (Swaminathan et al., 2012). However, this framework genetic map had a modest marker density which limited the usefulness for QTL mapping and marker assisted selection. In 2012, Ma et al. used Genotyping-By-Sequencing (GBS) to create a genetic map that identified 19 linkage groups based on 3745 markers (Ma et al., 2012). Since marker data from this research was proprietary, Liu et al. created a new high-density map with accessible SNP markers. Their map was based on 3182 SNPs, with an average intermarker spacing of 0.8 cM and 0.9 cM for respectively the female and the male map (Figure 2 (Liu et al., 2016).

More recently, Van der Weijde et al. (2017) published a novel genetic map based on a GBS approach. Their male map consisted of 242 SNP markers and had an average intermarker spacing of 8.0 cM. Their female map consisted of 322 SNP markers and had an average inter-marker spacing of 6.7 cM (van der Weijde et al., 2017b). The construction of the abovementioned maps has been facilitated by anchoring the DNA sequences and the linkage groups to the *Sorghum bicolor* genome. Furthermore, this anchoring facilitated a better interpretation of QTL mapping results and opened the possibility of comparative genomic studies. This resulted in the discovery of *M. sinensis* QTLs related to agronomic performance, combustion quality, biomass productivity, conversion efficiency characters and the zebra stripe phenotype (Gifford et al., 2015; Liu et al., 2016; Slavov et al., 2014; van der Weijde et al., 2017b). Some of these traits were found to have a very high number of QTLs, such as the discovery of 86 QTLs related to biomass composition (van der Weijde et al., 2017b). This indicates certain traits can have a high genetic complexity and a strong quantitative genetic control.

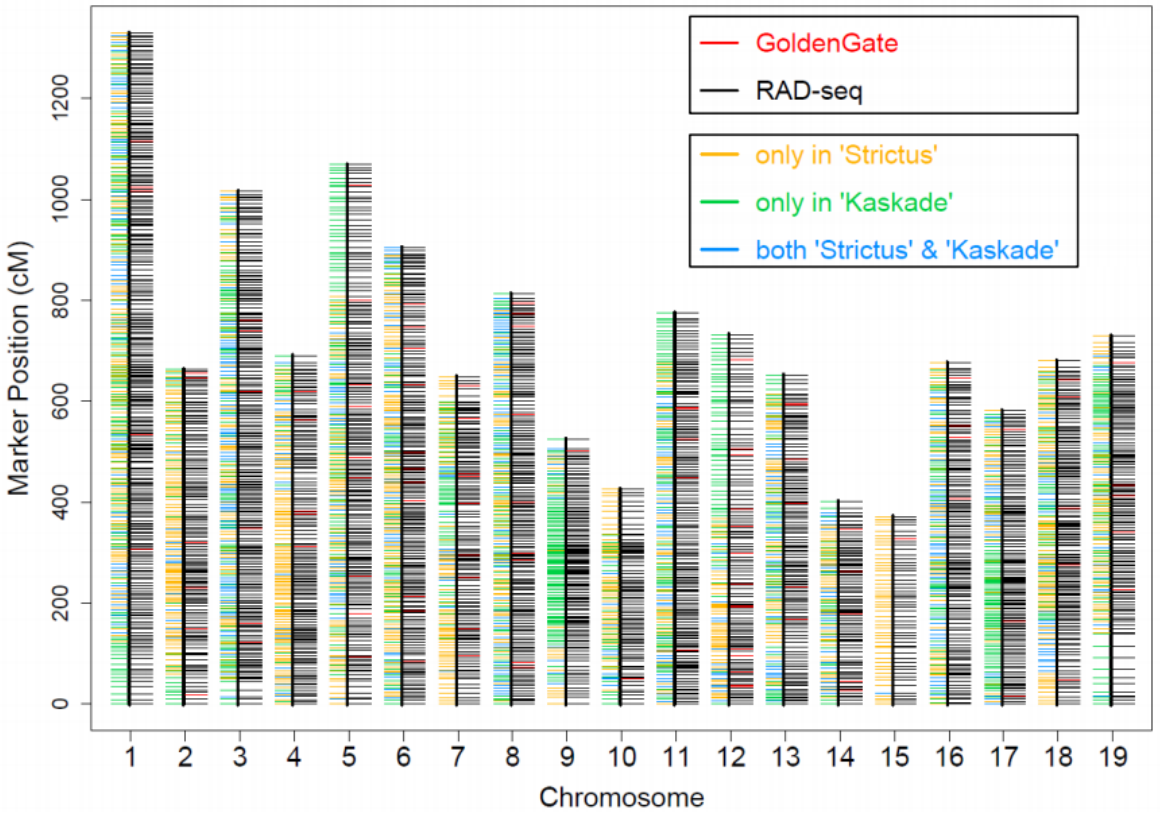


FIGURE 2: HIGH-DENSITY GENETIC MAP WITH ACCESSIBLE SNP MARKERS, ADAPTED FROM LIU ET AL. (2016) "GENETIC DISTANCE SHOWN ON THE LEFT IN CENTIMORGANS (CM). LINKAGE GROUP NUMBERS SHOWN AT THE BOTTOM (LINKAGE GROUP NUMBERS BASED ON THE GENETIC MAP FROM SWAMINATHAN ET AL. (2012). HORIZONTAL LINES REPRESENT ESTIMATED POSITIONS OF THE GENETIC MARKERS. MARKER TYPE IS SHOWN BY COLOUR (RED, GOLDENGATE SNPS; BLACK, RAD-SEQ SNPS; GOLD, HETEROZYGOUS IN M. SINENSIS 'STRICTUS' ONLY; GREEN, HETEROZYGOUS IN M. SINENSIS 'KASKADE' ONLY; BLUE, HETEROZYGOUS IN BOTH M. SINENSIS 'STRICTUS' AND M. SINENSIS "KASKADE')."

1.5 GENOMIC SELECTION

1.5.1 NEED

The emergence of next generation sequencing (NGS) techniques has led to the development of fast, cheap, high-throughput genome-wide SNP genotyping platforms. The development of methods such as restriction site associated DNA sequencing (RadSeq) and genotyping-by-sequencing (GBS) facilitated an even more efficient marker discovery. Both methods have been used for SNP discovery and map construction in *M. sinensis*, as described in section 1.4. On the whole, these methods contributed to the identification of over 10,000 QTLs in 12 mayor crops. However, from those thousands of QTLs, only a few have been actually used for marker assisted breeding programs (Bassi et al., 2016; Bernardo, 2008). A possible explanation is that, when dealing with complex traits, single markers often capture a too small part of the genetic variation for being useful in marker assisted breeding. An example of a complex trait is cell wall composition, for which plants devote approximately 10% of their genome, which is around 2500 genes (Carpita and McCann, 2008). In addition to these difficulties related to the complexity of traits, *M. sinensis* also has a large, complex genome that has undergone genome-wide duplication, which complicates interpretation of genotypic data and makes the search for informative SNPs and QTLs challenging (Ma et al., 2012).

1.5.2 THEORY

To make better use of the (increasing) availability of high-density, genome-wide markers a concept was proposed to use all genome-wide markers in a model to predict (complex) plant traits as the sum of all (minor) genetic effects. This approach that has been called genomic selection (Meuwissen et al., 2001). Genomic selection has a potential high advantage when breeding for complex traits that are regulated by multiple QTLs, such as yield and cell wall composition (Poland et al., 2012).

To perform genomic selection (GS), a population will be both phenotyped and genotyped, which will be referred to as the training population (TP). This training population will be used to create and train a statistical model that associates allelic data with phenotypic traits. The model then is used to predict a genomic estimated breeding value (GEBV) of non-phenotyped individuals, which are referred to as the breeding population (BP). The GEBV is calculated as the combination of useful loci that are in the genome of the individual, and gives an estimation for the usefulness of an individual as a breeding parent. Individuals with predicted superior phenotypes can be selected as parents in a very early stage and the duration of the breeding cycle is reduced drastically because no phenotyping is needed.

Parallel to the breeding program, the effectiveness of the prediction model needs to be assessed. This is done with a set of individuals that are genotyped and phenotyped, which is referred to as the validation population (VP). The GEBV for individuals of this population is calculated and its correlation to the actual observed phenotypic value indicates the accuracy of the model. Phenotypic information from the VP can be used to update and re-calibrate the model. Assessment of the effectiveness of the model using real data can be performed in three methods (Sallam et al., 2015).

Subset validation: The TP and VP are from the same set of lines and cross validation (e.g. 100 times 10-fold cross validation) is used. In this method, TP and VP are from the same environment, G x E interaction is not taken into account, so the accuracy is upwardly biased.

Interset validation: The TP and VP are predefined sets, and could be either environmentally defined sets (i.e. TP and VP have same genotype but have different environment) or chronologically defined (i.e. TP is an older line and is used to predict VP, which is either in the same or different environment)

Progeny validation:The TP includes parents (or grandparents) from the VP (which indirectly means they are in different environments)

1.5.3 EXPERIENCES

Genomic selection is based on the estimation of the effect of a given marker on the phenotype of an individual. There are several statistical models that can be used to calculate estimations of the marker effects. The difference between models is based on their assumptions when treating the variance of complex traits. Heslot et al., compared 11 genomic selection models on 8 different datasets and reached the conclusion that most models had a very similar accuracy for the given traits (Heslot et al., 2012).

Since genomic selection was described in 2001, many studies and reviews have been published to describe the effectiveness and discuss the prospects in plant breeding. Comparison of GS research in multiple crops and for several traits showed a very high potential for GS usage in plant breeding (Bhat et al., 2016). There are several factors that can influence the prediction accuracy of a GS model. Ten factors and their effect are described in Table 1.

TABLE 1: FACTORS THAT CAN INFLUENCE PREDICTION ACCURACY AND THEIR EFFECT (BASSI ET AL., 2016; BHAT ET AL., 2016; POLAND ET AL., 2012; SALLAM ET AL., 2015)

Factor	Effect
Size of TP	A bigger training population results in a higher prediction accuracy
Composition TP	More relatedness between TP and BP results in a higher prediction accuracy A higher population structure results in a lower prediction accuracy
H^2 trait	More heritability of a trait results in higher prediction efficiency More GxE interaction results in a lower prediction efficiency
Genetic architecture trait	Complex trait with high number of loci results in a lower prediction efficiency
Statistical model	Different statistical models result in different prediction accuracies
Validation method	Only validating within a similar environment (e.g. subset validation) will result in overestimation of H^2 and therefore overestimation of prediction efficiency
Marker type	Markers should be equally distributed over the genome. Both GBS and RadSeq have been proven to be effective for GS.
Marker density	The density should be high enough so each QTL will be in LD with a marker If multiple markers are linked to a QTL, its effect on a trait will be overestimated
Filtering of markers	Choice of the minimal minor allele frequency (MAF) and maximal missing data frequency influence the amount of (useful) markers for the GS model
Genotype imputation method	Missing marker data can be imputed in several ways (e.g. heterozygous value; mean imputation) and this can affect the prediction of the marker effect

To date, only one construction of a genomic selection model for *M. sinensis* has been reported (Slavov et al., 2014). In this research, RadSeq genotyping was used on a population of 138 genotypes and 20,000 SNPs were discovered. A model was constructed for 17 phenology, biomass and cell wall traits. Slavov et al. reported prediction accuracies ranging between 0.05 (dry matter) and 0.95 (moisture), with an average of 0.57 for the 17 studied traits. These results indicate the potential of genomic selection for miscanthus breeding programs.

Since genotyping performances have continued to improve, new predictions based on a higher density of SNPs are expected to have improved prediction accuracies. As described in section 1.3, the Wageningen UR miscanthus collection has a high variance in morphological and biomass quality traits that are important for the purpose of a biobased economy. Traits such as the saccharification efficiency can be predicted, which haven't been modelled before in GS studies in miscanthus.
A goal for breeding programs is to breed for new varieties that have a combination of several preferred characteristics, such as high yield, high conversion efficiencies and an optimal flowering time. Some of these traits are complex and can consist of dozens of QTLs. Furthermore, it takes two years before cell wall composition (i.e. biomass quality) and biomass production are predictive of the yield at full maturity. For an even more reliable prediction for these, and especially other traits, screening has to be performed after at least the third year (Arnoult et al., 2015; van der Weijde et al., 2017a). Because of these reasons, genomic selection has the potential to be a useful, cost- and time-efficient tool for the Wageningen miscanthus breeding program.

2 OBJECTIVES

The goal of the Wageningen UR miscanthus breeding program is to breed for a lignocellulosic crop that is better suited for the biobased economy. There is a need for novel, improved varieties with a higher morphological and biochemical quality. This implies breeding for low biomass recalcitrance and high conversation efficiencies. This reduces the processing costs, which is a limiting factor for the production of miscanthus on a larger scale.

This research is performed parallel to other research projects, which have as an overall goal to obtain more insight into the phenotypic and genotypic diversity of the Wageningen UR miscanthus collection and to develop novel breeding tools to improve the Wageningen miscanthus breeding program. This research has a specific goal to construct a genomic selection model as a novel breeding tool for the Wageningen miscanthus breeding program. The research will have 3 specific objectives:

1. To obtain updated information on the degree of phenotypic diversity (both, in morphological and biochemical quality traits) of the Wageningen UR miscanthus collection

2. To perform the molecular characterization and obtain insights into extent of genetic diversity of the Wageningen UR miscanthus collection

3. To construct and validate a genomic selection model to predict GEBVs for various morphological and biochemical traits in M. sinensis

3 MATERIALS AND METHODS

3.1 PLANT MATERIAL

The experimental population in this research is the Wageningen UR miscanthus collection. This collection consists of 128 plots, containing 106 *M. sinensis* genotypes, 13 *M. sacchariflorus* genotypes, 5 *M. x giganteus* genotypes and 4 other hybrid genotypes. The accessions originate from various international gene banks and breeding programs and show a high amount of phenotypic diversity. The collection was planted in 2012 (Wageningen, the Netherlands). Each plot consists of 16 plants (i.e. a square of 4x4 plants).

3.2 GENOTYPING

DNA was isolated from random young leaves from the middle plants within the plot for 94 *M. sinensis* plants from the collection. RAD-SEQ sequencing was performed by BGI (Shenzen, Guangdong, China). Briefly, DNA samples were cut with the restriction enzyme EcoR1 and unique adapters were ligated to the ends of the DNA fragments. DNA samples were pooled, purified and multiplied (PCR). DNA was sequenced using Ilumia HiSeq X10/4000 systems. A total amount of 371.52 Gb clean data was generated in the project. The quality of the obtained marker data was lower than expected. Unfiltered SNP calling resulted in 88000 SNPs, but after a 50% call rate threshold only 2600 SNPs remained. When a minor allele frequency (MAF) threshold of at least 3 allele copies of the minor allele was applied, only 2000 SNPs remained.
M. sinensis is still an orphan crop with little genomic resources. Therefore it is possible that the RAD fragments did not fit into the standard SNP calling pipeline and results were limited. Near the end of this thesis, an early version of the *M. sinensis* genome was published (Phytozome, 2018). BGI redid SNP calling using the recently published reference genome. This resulted in identification of over 8.1 million SNPS, of which 7.0 million were located on chromosomes and 1.1 million on scaffolds. Unfortunately, since the dataset arrived near the end of this thesis it could not be used for the analyses in this report.

3.3 PHENOTYPING

The miscanthus breeding program has been characterized for phenotypic traits relevant to plant morphology and biochemistry for several years. Continuing this work, a total of 14 traits have been measured across 2017 (Table 2). The methods that were used are similar to the methods used by Boogers (2017) and van der Weijde (2017a). Briefly, in January the four middle plants from the 5th growing season were harvested, dried (60^{o}C) and weighed. Starting from July, the plants from the 6th growing season were checked thrice a week for the first flowering plant (i.e. the moment the first flowering head was opening) and subsequently for the date that half of plot had a flowering plant. In the first week of October the number of stems per plant (> 30 cm) were counted for the 4 middle plants of the plot. Subsequently, plot lodging was characterized by a categorical score of 1-3, where: 1 was no lodging, 2 slight lodging and 3 severe lodging. In week 42 and 43 of October a harvest of 3 randomly chosen flowering stems from each of the 4 middle plants was performed, cutting 3-4 cm above ground level (i.e. a duplicate harvest). Leaves and flowers were stripped from the stem and for all three the fresh weight was measured. For the stem, length, node number and internode diameter were measured (i.e. thinnest diameter of middle internode). Stems where cut in pieces of 3-4 cm. Stems, leaves and flowers were dried (60^{o}C, 48h) and dry weighs were measured. Dried stems were milled using a hammer mill with a 1 mm screen.
Cellulose, hemicellulose and lignin contents of the stem cell wall were estimated by determining NDF, ADF and ADL contents (in duplicate) using Ankom technology. Samples and buffers were prepared according to protocols developed by Ankom Technology (2017a; 2017b; 2017c). For step 8 and 9 in the ADL protocol, treatment was performed by placing 1L Schott flasks containing 24 bags covered with 72% H_2SO_4 in a shaker for 3 hours at 120rpm at a 45° angle.

Enzymatic saccharification efficiency was measured similar to the protocol described by Van der Weijde et al. (2017a). The preeminent deviation was the usage of an alternative enzymatic cocktail for digestion, since the original cocktail became unsuitable for independent research purposes. Briefly, samples were prepared by weighing 500 mg stem subsample in F57 Ankom filter bags (in triplicate). The biomass was washed with deionized water (3x 5min, ~60°C) for the removal of soluble sugars, after which the samples were folded and placed in 50ml Falcon tubes. An alkaline pre-treatment was performed (2 hours incubation in 2% NaOH, at 50°C at 160 rpm) after which the samples were thoroughly washed and neutralized. To determine saccharification efficiency the pre-treated biomass was hydrolysed by a 150 µl cellulase enzyme blend (1.2 g/ml; Sigma-Aldrich, Saint Louis, MO, US) and 15 µl endo-1,4-ß-xylanase M1 (1,700 U/mL; Megazyme, Bray, IE), for 48 hours in an incubator shaker at 50°C at 160 rpm. Reactions were carried out in 44 ml 0.1M sodium citrate buffer (pH = 4.8), containing 0.375 g/L sodium benzoate to prevent microbial contamination. After saccharification the enzymatic activity was stopped by incubation for 5.5 minutes at 99°C at 300 rpm. Samples were centrifuged and diluted 50x after which the sugar release was quantified using High-Performance Anion-Exchange Chromatography (HPAEC) analysed by a Dionex system equipped with a CarboPac PA1 column and a pulsed amperometric detector (Dionex, Sunnydale, CA).

TABLE 2: PHENOTYPIC TRAITS OF INTEREST, MEASURED IN 2016 AND 2017 (RESP. GROWING YEAR 5 AND 6)

Trait	Definition
Stem number	Mean number of flowering stems per plant (cm)
Stem length	Mean stem length (cm)
Internode diameter	Mean diameter of middle internode (m)
Internode length	Mean internode length per plant (cm)
Stem yield	Total dry weight of 12 flowering stems (g DW)
Leaf yield	Total dry weight of the leaves of 12 flowering plants (g DW)
Flower yield	Total dry weight of the flowers of 12 flowering plants (g DW)
Biomass yield	Total dry weight of the 4 middle plants per plot after harvest (g DW)
Initiation Flowering	Date of flowering initiation(Julian Day Number, JDN)
50% Flowering	Date on which 50% of the plants have a flowering stem (Julian Day Number, JDN)
Cellulose	Cellulose content, gravimetrically measured (% dry weight)
Hemicellulose	Hemicellulose content, gravimetrically measured (% dry weight)
Lignin	Lignin content, gravimetrically measured (% dry weight)
CelCon	Percentage of cellulose converted to glucose (% dry weight)

3.4 DATA ANALYSIS

Phenotypic data analyses have been performed using Genstat v18 (VSN_International, 2015). Missing data for flower yield and total biomass yield were estimated via their correlation to 'stem yield' and 'stem yield x stem number', respectively. Plants that did not flower were assigned the highest observed value. Outliers in the saccharification efficiency experiments were left out for analyses.

Summary statistics were calculated for all phenotypic data from the 4th and 5th growing season. ANOVAs were performed for each trait using the different growing season as blocking structure. Broad sense heritability's were estimated as

$$H^2 = \frac{V_{Genotype}}{V_{Phenotype}} = \frac{MS_{plot}}{MS_{plot}+MS_{residual}} \qquad \text{(Equation 1)}$$

Histograms of the phenotypic variance were generated using the R package ggplot2.

For each SNP, the sequence of the 82bp RAD fragment from which the SNP was called was known. Geneious v10.0.9 (Kearse et al., 2012) was used to perform a local BLAST of these sequences. First a BLAST database was built for a custom program from an early release of the Miscanthus genome (Phytozome, 2018). Afterwards, the sequences were BLASTed against this database via the 'Megablast' method, to obtain a 'query-centric alignment' with a 'matching region with annotations'. The output showed the best hits for each sequence on the genome, with information about the position (e.g. chromosome, hit start) and their match (e.g..% Pairwise Identity, E value and grade).

R base was used to perform a principal component analysis. Genomic prediction was performed using custom scripts, adapted from scripts developed by Mario Callus from the WUR Animal Breeding and Genetics department (Appendix 1). Additionally, genomic prediction was performed using the rrBLUP package (Endelman, 2011). Custom scripts used a leave-one-out cross validation whereas the rrBLUP package used a larger subset validation with a customizable amount of cycles. The default parameters for the custom scripts were 2596 SNPs with an SNP calling of > 0.5, without removal of markers below MAF threshold, using custom broad sense heritability estimates and a leave-1-out cross validation.

4 RESULTS AND DISCUSSION

4.1 THE WUR MISCANTHUS COLLECTION DISPLAYS A BROAD RANGE OF PHENOTYPIC DIVERSITY THAT IS HIGHLY HERITABLE

Since *M. sinensis* is still an orphan crop there is an urge to use novel breeding tools that can accelerate breeding efforts. These efforts should aim to design a crop varieties with favourable characteristics that can be adapted to the requirements of biorefineries. Therefore it is of high importance to establish diverse germplasm collections, characterizing their range of phenotypic diversity and identifying the genetic determinants thereof.
The WUR miscanthus collection has been phenotyped during the 4th and 5th growing season. All traits show a broad range of phenotypic diversity, with the coefficient of variation ranging between 6.8% for lignin and 63.0% for stem yield (Table 3). The highest variation is measured for yield related traits, which is promising since yield improvement is one of the primary objectives of international miscanthus breeding (Weijde et al., 2013). Although the coefficients of variation are generally lower for cell wall quality traits, there is still ample variation between genotypes. The observed variation in hemicellulose and lignin content was in accordance with previous reports (Allison et al., 2011; van der Weijde, 2016). Cellulose content exceeded the reported range of 26-49%. Similarly, cellulose conversion efficiency exceeded the reported range of 32-50%. These findings can be caused by a difference in phenotyping, environmental effects, the availability of novel germplasm in our experimental population or any combination of these.

TABLE 3: AVERAGES OF MORPHOLOGICAL AND BIOCHEMICAL TRAITS (GROWING SEASON 2016 AND 2017)

Summary statistics	Mean	Minimum	Maximum	CoV
Stem number	68	12	206	46.7%
Stem length	183	66	298	25.9%
Internode diameter	4.81	2.10	7.66	22.7%
Internode length	19.1	9.7	34.7	18.0%
Stem yield	142	18	456	63.0%
Leaf yield	84.0	6.4	265.2	57.5%
Flower yield	16.4	2.9	42.0	51.9%
Biomass yield	1904	213	4464	45.5%
Initiation Flowering	233	183	282	10.9%
50% Flowering	242	186	283	11.4%
Cellulose (%)	50%	40%	57%	6.8%
Hemicellulose (%)	30%	22%	35%	8.4%
Lignin (%)	10%	5%	15%	19.5%
CelCon(%)	50%	30%	79%	26.9%

Most morphology-related traits tend to follow a normal distribution, with the distributions of both years overlapping each other (Figure 3). Only one trait, 50% flowering, follows a clear different pattern that resembles a bimodal distribution. This flowering behaviour can be explained by a significant correlation between 50% flowering and the maximum temperature per day (Bogers, 2017). Biomass yield is normally distributed, but the 5th growing season has relative higher yields. Similarly, traits related to cell wall quality all tend to follow a normal distribution, but the distribution mean of both years differ (Figure 4). For the 5th growing season, plants had on average higher cellulose, higher lignin but lower hemicellulose relative to the previous year. The phenotypic differences in biomass yield and cell wall composition between both years can be caused by several factors.

Genetic, agronomical (e.g. harvest date), climatic and other environmental factors have been shown to influence biomass production (Arnoult and Brancourt-Hulmel, 2015) and cell wall composition of *Miscanthus* (Allison et al., 2011; Golfier P, 2016). An alternative explanation is that the plants had not fully matured and traits were not stable yet. Comparing the similar distributions per year for morphology-related traits to the shifted distributions for the cell wall quality related traits reveals that, although on a morphological level the plants look similar the changes on the biochemical level can be significant.

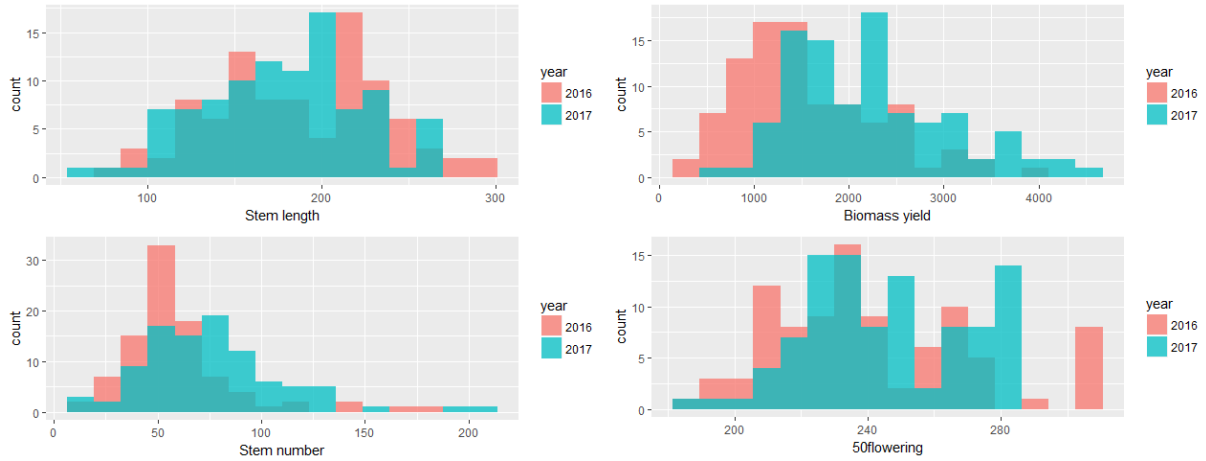


FIGURE 3: EXTENT OF PHENOTYPIC VARIATION IN IMPORTANT MORPHOLOGICAL TRAITS IN GROWING SEASON 2016 AND 2017

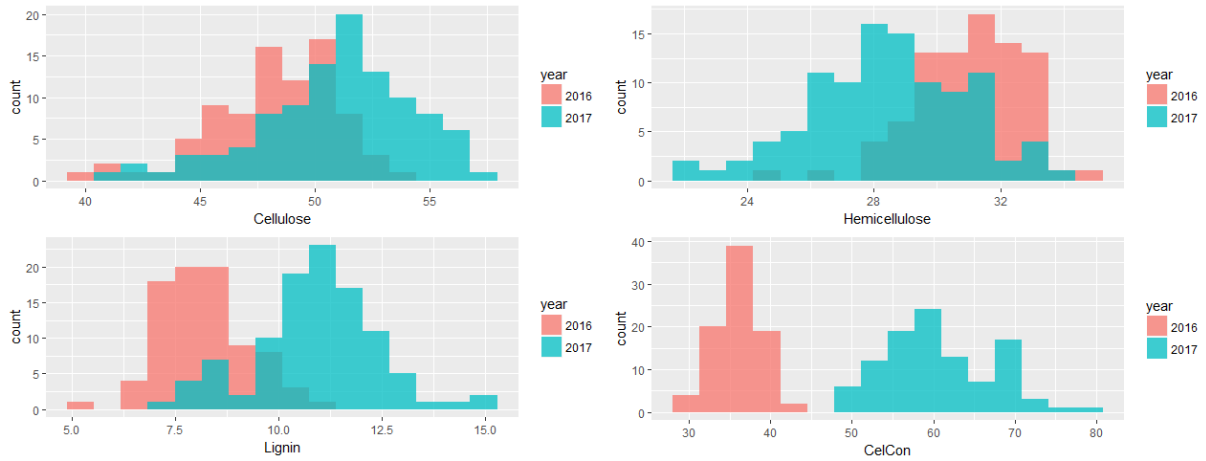


FIGURE 4: EXTENT OF PHENOTYPIC VARIATION OF CELL WALL CHARACTERISTICS IN GROWING SEASON 2015 AND 2016

The two distributions for cellulose conversion efficiency have distinct means, which is a result of the usage of two different enzymatic cocktails. This resulted in a high range of values that explains the high coefficient of variance for CelCon compared to the other cell wall quality traits (Table 3). Both enzymatic cocktails have a characteristic enzymatic performance and genotypes could perform differently between the different treatments. Therefore one might argument that the different measurements can be considered different traits. However, although the broad sense heritability (H^2) estimates are lowest of all traits they still show that a considerable amount of the phenotypic variation can be explained by the variation in the genotypes (Table 4). Future experiments will show whether the H^2-estimate for CelCon was realistic or if it was an underestimation caused by the measurements differences.

The different ranges in the enzymatic saccharification experiment address the need for robust phenotyping protocols. Robust phenotyping protocols will improve phenotyping precision and therefore improve the power of breeding efforts. At a more general level, robust and uniform phenotyping protocols will facilitate the comparative value of data across different years, different environments and even different studies (Arnoult and Brancourt-Hulmel, 2015). Another phenotyping protocol that can be improved within this study is the separation of leaves and stems after harvest. Incomplete leaf removal will result in overestimations of stem yield. Moreover, leaf and stem can have different biochemical composition that may be masked when not separated completely (da Costa et al., 2014).

TABLE 4: BROAD SENSE HERITABILITY ESTIMATES

H^2:	
Stem number	0.89
Stem length	0.94
Internode diameter	0.92
Internode length	0.84
Stem yield	0.89
Leaf yield	0.93
Flower yield	0.85
Biomass yield	0.84
Initiation Flowering	0.89
50% Flowering	0.96
Cellulose (%)	0.96
Hemicellulose (%)	0.91
Lignin (%)	0.83
CelCon (%)	0.77

The high degree of phenotypic variation that is observed from Table 3, Figure 3 and Figure 4 can be largely attributed to variation in the genotypes (Table 4). The H^2-estimates have an average of 0.89 and range from 0.77 for CelCon up to 0.96 for Cellulose and 50% flowering. In practice, these H^2-estimates show that a trait is likely to behave similar in the next growing season. Arnoult and Brancourt-Hulmel (2015) reviewed several studies on miscanthus and concluded that all traits had a high contribution of the genotype to the phenotypic variability. Nevertheless, the H^2-estimates observed in this study are higher expected. Slavov et al. (2014) reported an average H^2 of 0.64, with heritabilities for cellulose, hemicellulose and lignin content of 0.79, 0.60 and 0.66, respectively. These differences in trait heritability are most likely caused by differences in experimental design. The study population of Slavov et al. followed a randomized complete block design with four blocks and one replicate per genotype per block, with the plants grown at 1.5x 1.5 m spacing. The WUR miscanthus collection has no biological replicates but each plot contains 16 plants to avoid border effects. In this study the growing season was used as blocking structure to estimate heritabilities, whereas Slavov et al. used biological replicates as blocking structure. Whether variation between locations in the field, or variation between years will result in a bigger environmental variation is dependent on both the specific conditions each year and on the specific conditions across the experimental field. Regarding research standards on experimental designs, it would be advised to have biological replicates in a randomized design to be able to cancel out site-specific environmental effects. Nevertheless, the relatively high heritabilities observed in this study show the significance of avoiding border effects to reduce environmental variation and its positive effect on heritabilities.

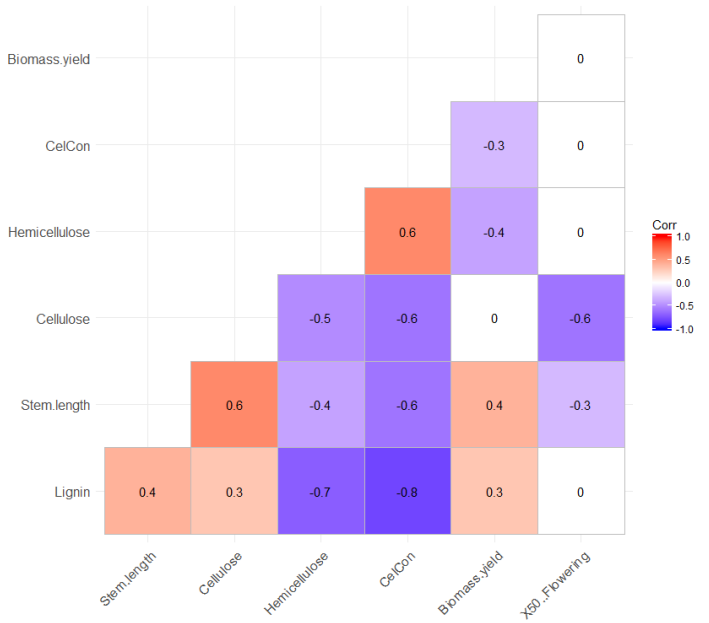


FIGURE 5: PHENOTYPIC RELATIONSHIP MATRIX FOR IMPORTANT BIOBASED-RELATED TRAITS, INSIGNIFICANT RESULTS HAVE BEEN SET TO ZERO. A FULL PAGE MATRIX OF ALL TRAITS IS SHOWN IN APPENDIX 2

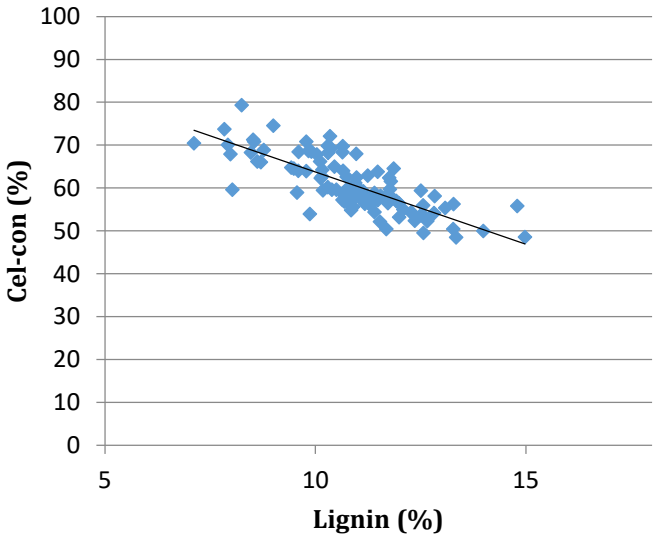


FIGURE 6: CORRELATION BETWEEN LIGNIN CONTENT AND CELLULOSE CONVERSION EFFICIENCY

Strong correlations were observed between traits, with as strongest a negative correlation between lignin content and cellulose conversion efficiency (Figure 5,Figure 6). This negative correlation is widely known in literature, and it is why reduction of lignin content has been targeted as another of the major breeding goals for Miscanthus (Arnoult and Brancourt-Hulmel, 2015; Golfier P, 2016; Grabber, 2005; Weijde et al., 2013). Another strong negative correlation is observed between lignin content and hemicellulose content. Van der Weijde et al. (2017c) have suggested that hemicelluloses and lignin have similar functions regarding cell wall rigidity. Hence, reductions in lignin content can be compensated by increase of hemicellulose content and therefore do not necessarily lead to lower cell wall rigidity and plant fitness. This theory was supported by a positive correlation between hemicellulose content and cellulose conversion, which is also observed in this study.

In conclusion, all traits characterized within the WUR miscanthus collection show a broad range of phenotypic diversity with a high amount of variation that can be used in breeding programs. Heritability estimates have shown that this phenotypic variation for all traits can be largely attributed to the genotypic determinants. This shows even better the high quality and possibilities the WUR miscanthus collection possesses for breeding novel improved miscanthus accessions.

4.2 MOLECULAR ANALYSES SHED LIGHT ON GENETIC RELATIONSHIPS, BUT ARE CHALLENGED BY A LARGE, REPEAT-RICH GENOME

The previous section has shown the WUR miscanthus collection has a high phenotypic diversity. To facilitate breeding efforts and usage of molecular breeding tools it is valuable to have insight into the genetic diversity of the experimental population. Principal component analysis on 2600 SNPs shows that two principal components are already well able to discriminate between the accessions (Figure 7). One dense cluster of genotypes is located at the right-upper corner of the plot. Another group is located in the left upper corner (i.e. which shows clear distinction on PC1). Most of the remaining genotypes are distributed over PC2, mostly on right half of PC1. The two first principal components (PCs) explain a relatively large proportion (38.2%) of the variation. This can be caused by the low amount of markers that was used (2600 SNPs). When 22,000 and 80,000 SNPs were used for analysis, the total variation explained by the first two PCs changed to 26.2% and 17.3%, respectively. This means that the addition included markers that were uncorrelated to the initial markers. However, because the extra SNPs that were added had a very low call rate with up to 90% missing data, these analyses should be repeated with a larger amount of markers of higher quality.

Interestingly, the PCA shows multiple occurrences of genotypes originating from the same research project grouping together. One clear example is the group of 111 up to 128 on the left-upper corner that consist mostly of individuals from the '1997' and '1998' groups. Within the experimental field, individuals from similar research projects have been planted together in a consecutive order. These findings indicate that several of these consecutive plots have relatively high genetic similarity. It is known that the environment can have a significant effect on the *M. sinensis* phenotype (Allison et al., 2011; Arnoult and Brancourt-Hulmel, 2015; Golfier P, 2016). Physically close positioning of genetically close accessions can lead to an inability to discriminate between genotypic or environmental effects. To illustrate this a subset including the individuals from the '1997' and '1998' groups was compared to the remaining accessions. Both subsets show a different distribution for stem length and cellulose content (Figure 8). Certain molecular breeding tools such as genomic prediction use genomic relationships to predict phenotypes. If genetically similar individuals experience similar environmental effects predictions will be more accurate, which would be an overestimation. Incorporating a randomized design into a future experimental field could overcome these problems.

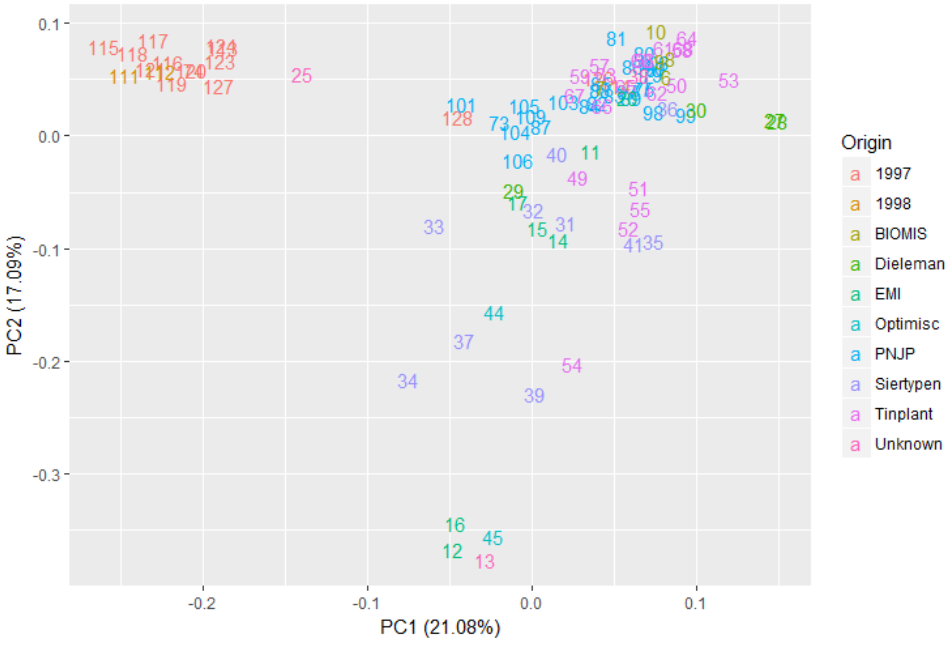


FIGURE 7: PRINCIPAL COMPONENT ANALYSIS OF THE GENOMIC RELATIONSHIPS, LABELS ARE PLOT NUMBERS, FULL PAGE IMAGE IN

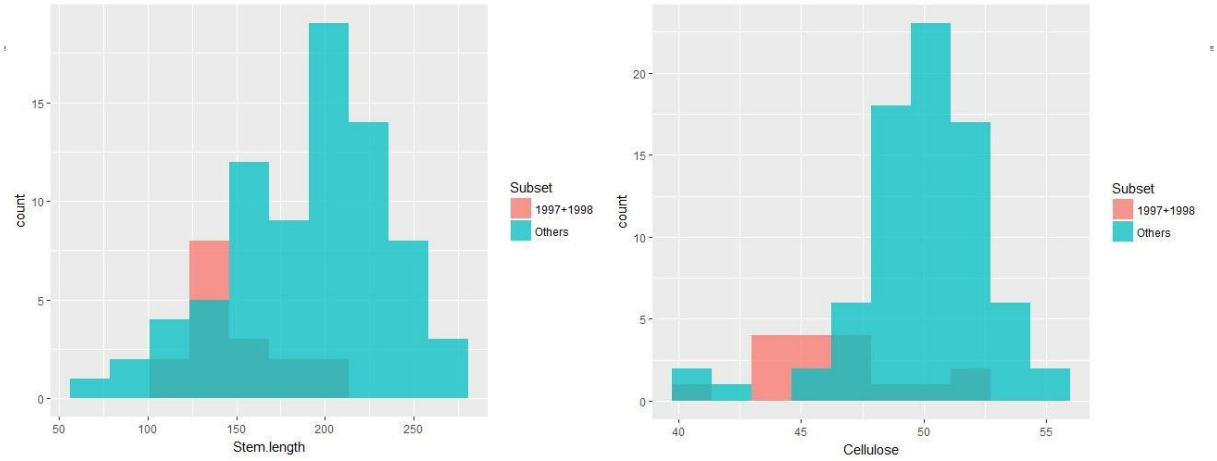


FIGURE 8: PHENOTYPIC DISTRIBUTION OF SUBSET '1997' AND '1998' COMPARED TO THE REMAINING ACCESSIONS FOR THE TRAITS STEM LENGTH AND CELLULOSE CONTENT

Alignment of the 82bp RAD fragments against the recently published reference genome showed a relatively balanced genome coverage (Figure 9). Contrary to literature, no bias for putative centromeric regions was detected (Slavov et al., 2014). Slavov et al. performed RAD-sequencing with a methylation sensitive restriction enzyme and already suggested that usage of multiple restriction enzymes with varying sensitivities to methylation could mitigate the bias in SNP detection. In this study it is shown that usage of the methylation insensitive EcoRI resulted in an improved genome coverage. However, the importance of a balanced genome coverage is debateable. For *sorghum bicolor* a high diversity was observed in gene-rich euchromatic regions. Heterochromatic pericentromeric regions showed low genetic diversity and low recombination rates (Evans et al., 2013). Since the *Miscanthus* genome is similar to the *Sorghum bicolor* genome, similar patterns in gene density between euchromatic and pericentromeric heterochromatic regions can be expected. Therefore a lower SNP coverage in pericentromeric regions might not be an issue for molecular studies. In *Zea mays*, trait-associated SNPS were found to be particularly enriched in nongenetic regions within 5kb upstream of genes. These regions are often related to regulation of gene expression (Yu et al., 2012). However, these intergenetic regions are often targeted for methylation, which in *Zea mays* can be up to 50% of cytosine being methylated (Suzuki and Bird, 2008). If the genetic organisation within the *Miscanthus* genome is comparable to *Z. mays,* usage of a methylation sensitive enzyme may prevent trait-associated SNPs from being detected. Considering these findings, it has yet to be determined which restriction enzymes would be optimal to obtain a markers set that yields most information for molecular studies.

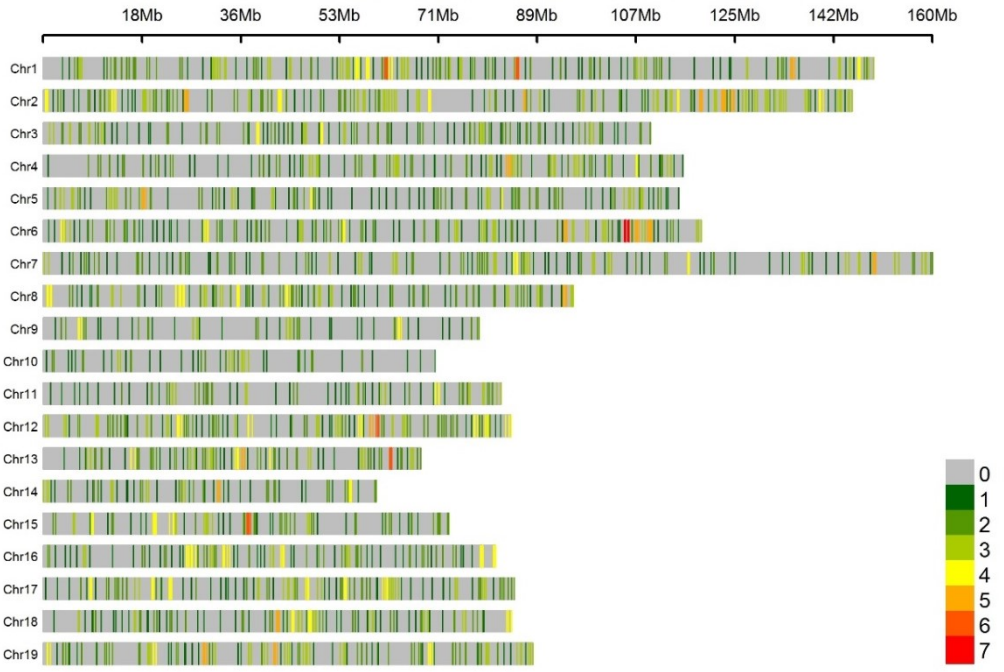


FIGURE 9: DISTRIBUTION OF SNPS FOR ALL 19 CHROMOSOMES OF M. SINENSIS

From BLASTing the RAD-fragments it was observed that, in the majority of the cases, not only the best hit but also successive hits had a high similarity to regions of the genome (Table 5; Table 6). It is known *M. sinensis* evolved from a genome wide duplication in its progenitor, which was closely related to a *Sorghum* ancestor (Ma et al., 2012). However, a single genome duplication would not explain the observed high similarities. Miscanthus shares a similar genome structure to maize, which also had a genome duplication in the recent past (Ma et al., 2012; Schnable et al., 2012). The large genome size of maize has been explained as the result of transposon amplification. Long terminal repeat (LTR) retrotransposons can constitute to 75% of the maize genome (Schnable et al., 2009). In comparison, the sorghum genome contains approximately 55% LTR-retrotransposons, whereas the relatively small rice genome contains 26% LTR-retrotransposons. For sorghum, a repeat content of approximately 61% was reported (Paterson et al., 2009). Aside from a large number of LTR retrotransposons in Sorghum, copy number variations were detected in several thousand genes. Some of these genes could be related to basic biological functions and even bio-energy related traits (Paterson et al., 2009). These findings shed light onto the similarity within genomes of grasses and can explain the high amount of highly similar BLAST hits. In future, databases such as Grassius that integrate gene regulatory information for maize, rice and sorghum and will facilitate Miscanthus genetic research and improve knowledge of regulatory genes in Miscanthus (Jakob et al., 2009).

TABLE 5: FIRST 4 BLAST HITS FOR RAD-FRAGMENT 'RECORD_2233'

HIT	SEQUENCE NAME	SEQUENCE LENGTH	HIT START	HIT END	% PAIRWISE IDENTITY
1	Chr14	82	27,104,280	27,104,361	99.40%
2	Chr13	82	29,143,087	29,143,168	98.20%
3	Chr01	82	13,892,769	13,892,688	98.20%
4	Chr13	82	29,719,085	29,719,166	98.20%

TABLE 6: AVERAGES OF THE 4 BEST HITS FOR BLASTING THE RAD-FRAGMENTS AGAINST THE M. SINENSIS REFERENCE

HIT	AVERAGE (%) PAIRWISE IDENTITY	AVERAGE E VALUE	AVERAGE GRADE	TOTAL HITS
1	98.82%	4.00E-16	98.57%	2501
2	96.64%	4.00E-16	93.79%	1713
3	96.08%	4.00E-16	92.91%	866
4	96.18%	4.00E-16	93.95%	740

The genomic information obtained in this study gave a first insight into the genetic relationships within the experimental population and displayed a promising genome coverage. However, it was shown that genetic studies using the repeat-rich *M. sinensis* genome are challenging and the alignment results are not set in stone. Future genetic studies into the reference genome will provide a deeper understanding in the repeatability and homology of the *M. sinensis* genome.

4.3 EXPLORATORY FINDINGS ON GENOMIC PREDICTION ACCURACIES AND THE DETERMINING PARAMETERS THEREOF

It has been shown that the WUR miscanthus collection has a high degree of phenotypic variation, with high estimates of heritability. Sequencing results of 94 *M. sinensis* genotypes had a lower quality than expected but had a relatively balanced distribution across the genome and genotypes could be distinguished based on their genetic relationships. Therefore it was concluded that the WUR miscanthus collection has potential to be used in *M. sinensis* breeding programs. However, as discussed in the introduction, classical breeding strategies for *M. sinensis* breeding are slow and there is a strong urge for molecular genetic tools. Genomic Selection has been widely regarded as a promising crop breeding tool and after first tests for *M. sinensis* it was concluded that genomic selection could be immediately applied in breeding programs (Slavov et al., 2014).

In this study the potential of genomic selection for the WUR miscanthus collection was assessed. Briefly, genome estimated breeding values (GEBVs) of individuals were predicted based their genotypic relationships with other individuals. These GEBVs were correlated with the real phenotypic value to estimate prediction accuracies (Figure 10). Generally, prediction accuracies were high under default settings (Table 7). Prediction accuracies had an average of 0.51, ranging from 0.32 for lignin content to 0.72 for stem length. The average prediction accuracy of 0.51 in this study is lower than the average prediction accuracy of 0.57 reported by Slavov et al. This might be because different traits were measured, but it is most likely caused by the larger amount of better quality markers Slavov et al used.

The effect of enlarging the set of markers to 22,500 or 82,000 SNPs was trait-dependent. The effect could be either positive (e.g. stem number), negative (e.g. lignin) or little effect at all (e.g. stem length). Added markers can contain new useful information that can be used for a better prediction. On the other hand, adding low quality markers can also add a 'noise' that reduces predictive abilities (M. Callus, personal communication). This addresses the importance of high quality marker data with a high call rate. Removal of SNPs that were below M.A.F-threshold had no effect on prediction accuracies. This is in accordance with other studies. Nevertheless, in large datasets removal of SNPs below M.A.F.- threshold is arbitrary to correct for mistakes in SNP calling (M. Callus, personal communication). Using a smaller training population (88 genotypes) to predict a bigger validation population (6 genotypes) resulted in lower prediction accuracies. This was expected since there is less information in the training population. For this procedure the 88 genotypes were randomly selected from the total population and this was iterated for 100 and 500 cycles. Given that the prediction accuracies could change up to 8% between the different iterations indicates that there should be sufficient cycles if the model validation is assessed in this way.

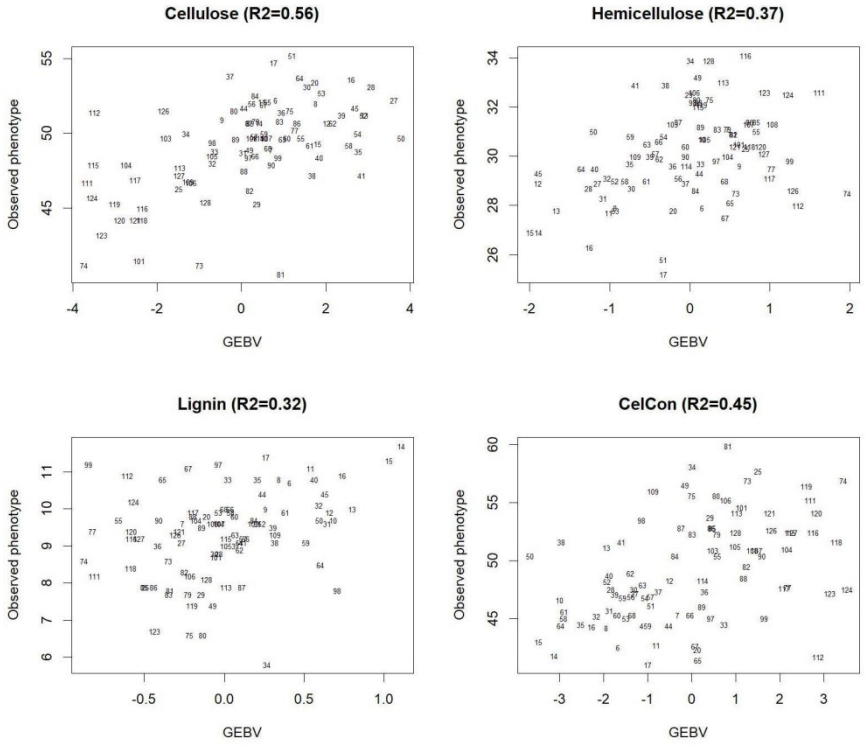


FIGURE 10: PLOTS FOR THE REGRESSION OF GEBV AND OBSERVED PHENOTYPE FOR MOST IMPORTANT BIOCHEMICAL TRAITS

TABLE 7: MEAN PREDICTION ACCURACIES ACROSS DIFFERENT GENOMIC PREDICTION MODELS

	2600 SNPs (default)	22500 SNPs	82000 SNPs	2000 SNPs (– M.A.F.)	88 geno 100 cycles	88 geno 500 cycles
Stem number	0.50	0.58	0.57	0.50	0.38	0.35
Stem length	0.72	0.71	0.71	0.72	0.70	0.69
Internode diameter	0.54	0.62	0.61	0.54	0.63	0.56
Internode length	0.48	0.56	0.55	0.48	0.42	0.34
Stem yield	0.64	0.67	0.67	0.64	0.68	0.60
Leaf yield	0.58	0.62	0.61	0.58	0.59	0.60
Flower yield	0.40	0.53	0.52	0.40	0.30	0.30
Biomass yield	0.45	0.35	0.34	0.45	0.42	0.38
Initiation Flowering	0.56	0.62	0.59	0.56	0.50	0.53
50% Flowering	0.57	0.63	0.61	0.57	0.58	0.55
Cellulose (%)	0.56	0.58	0.58	0.56	0.53	0.53
Hemicellulose (%)	0.37	0.40	0.44	0.36	0.41	0.40
Lignin (%)	0.32	0.25	0.24	0.32	0.29	0.32
CelCon (%)	0.45	0.43	0.44	0.45	0.44	0.45
Average	0.51	0.54	0.53	0.51	0.49	0.47

Prediction of traits based on the average values from both growing seasons was compared to prediction based on the individual growing seasons (Table 8). In 2017, certain morphology-related traits had 10-15% lower prediction accuracy relative to 2016 (e.g. stem number, stem length). On the other hand, cell wall quality traits could be predicted 10-15% better relative to 2016. This indicates that differences in phenotypes per year can have a considerable effect on the prediction accuracy. As discussed in chapter 4.1, more robust and uniform phenotyping protocols and randomized experimental designs could help to overcome these problems. Assuming no differences in phenotyping, trait averages of multiple years would be the most realistic predictors (M. Callus, personal communication). Another explanation of the differences in accuracies between years could be the presence of influential data points in the prediction accuracy regression. Histograms of the phenotypic distribution showed that in 2016 the number of stems per plants followed a smaller normal distribution compared to 2017 (Figure 3). Both years have a few high performing genotypes, which are influential data point in the regression that affect the slope of the regression and therefore the prediction accuracies (Figure 11). This finding addresses the usefulness of correlation plots as a quick visual data check, and a means to obtain insight into which are the best and worst predicted genotypes.

TABLE 8: MEAN PREDICTION ACCURACIES OF GENOMIC PREDICTION MODELS (1'FOLD'VALIDATION) USING PHENOTYPIC DATA FROM DIFFERENT YEARS

	Both	2016	2017
Stem number	0.50	0.53	0.38
Stem length	0.72	0.73	0.68
Internode diameter	0.54	0.53	0.50
Internode length	0.48	0.47	0.44
Stem yield	0.64	0.61	0.72
Leaf yield	0.58	0.52	0.61
Flower yield	0.40	0.36	0.47
Biomass yield	0.45	0.23	0.56
Initiation Flowering	0.56	0.63	0.57
50% Flowering	0.57	0.62	0.68
Cellulose (%)	0.56	0.54	0.69
Hemicellulose (%)	0.37	0.35	0.52
Lignin (%)	0.32	0.38	0.49
CelCon (%)	0.45	0.55	0.62
Average	0.51	0.50	0.57

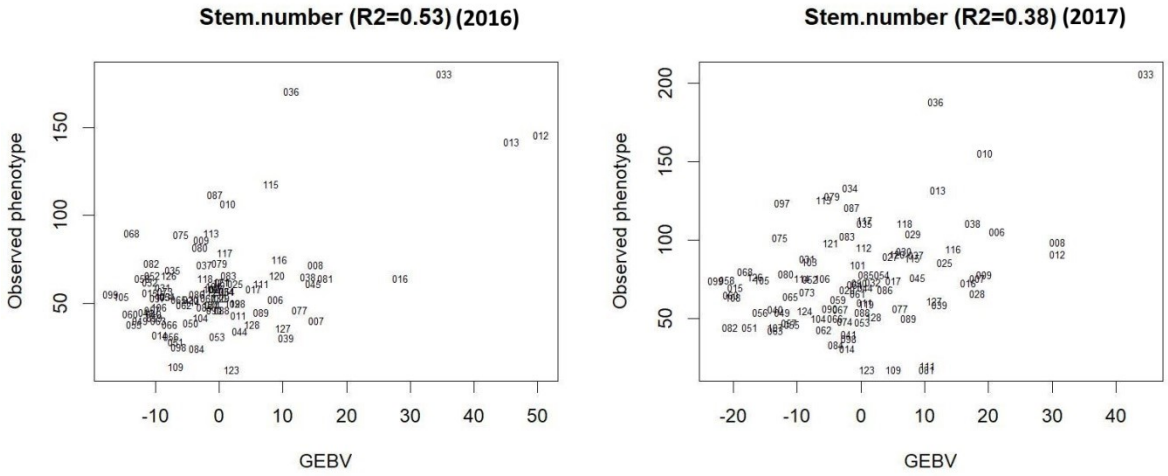


FIGURE 11: REGRESSION PLOTS FOR STEM NUMBER FOR THE YEARS 2016 AND 2017

Average prediction accuracies improved slightly when our own heritability estimates were used instead of the arbitrary value that was initially used in the model (Table 9). An increase in heritability is indeed expected to be beneficial because the component of accuracy due to genetic relationships will gain in importance (Jannink et al., 2010). This finding adds to discussion in chapter 4.1 regarding the importance of accurate heritability estimation. For genomic prediction it is especially important to have accurate H^2-estimates, since it is only possible to predict the part of the phenotype that is due the genotype. In other words, H^2 is the upper limit genomic prediction can reach.

TABLE 9: MEAN PREDICTION ACCURACIES FOR DIFFERENT HERITABILITY SETTINGS. FOR THE THIRD COLUMN OUR OWN BROAD-SENSE HERITABILITY ESTIMATES WERE USED.

	H^2=0.3	H^2
Stem number	0.39	0.50
Stem length	0.68	0.72
Internode diameter	0.46	0.54
Internode length	0.46	0.48
Stem yield	0.59	0.64
Leaf yield	0.51	0.58
Flower yield	0.33	0.40
Biomass yield	0.36	0.45
Initiation Flowering	0.52	0.56
50% Flowering	0.55	0.57
Cellulose (%)	0.60	0.56
Hemicellulose (%)	0.36	0.37
Lignin (%)	0.29	0.32
CelCon (%)	0.43	0.45
Average	**0.47**	**0.51**

Manhattan plots showed markers with high effect located on all chromosomes (Figure 12). However, in this study only 2269 markers with genomic positional information were available. For the *M. sinensis* draft genome 67,789 loci containing protein-coding transcripts were reported (Phytozome, 2018). This means there was roughly one marker per 30 genes. In future, experiments with a higher density of high quality markers are expected to give more insight into important regulatory regions on the genome. On a chromosome level, patterns of physically close markers with similar (high) SNP effects might be observed. The potential information about high-effect SNPs should be integrated with GWAS and QTL analysis results to detect similar patterns in order to obtain improved biological interpretations.

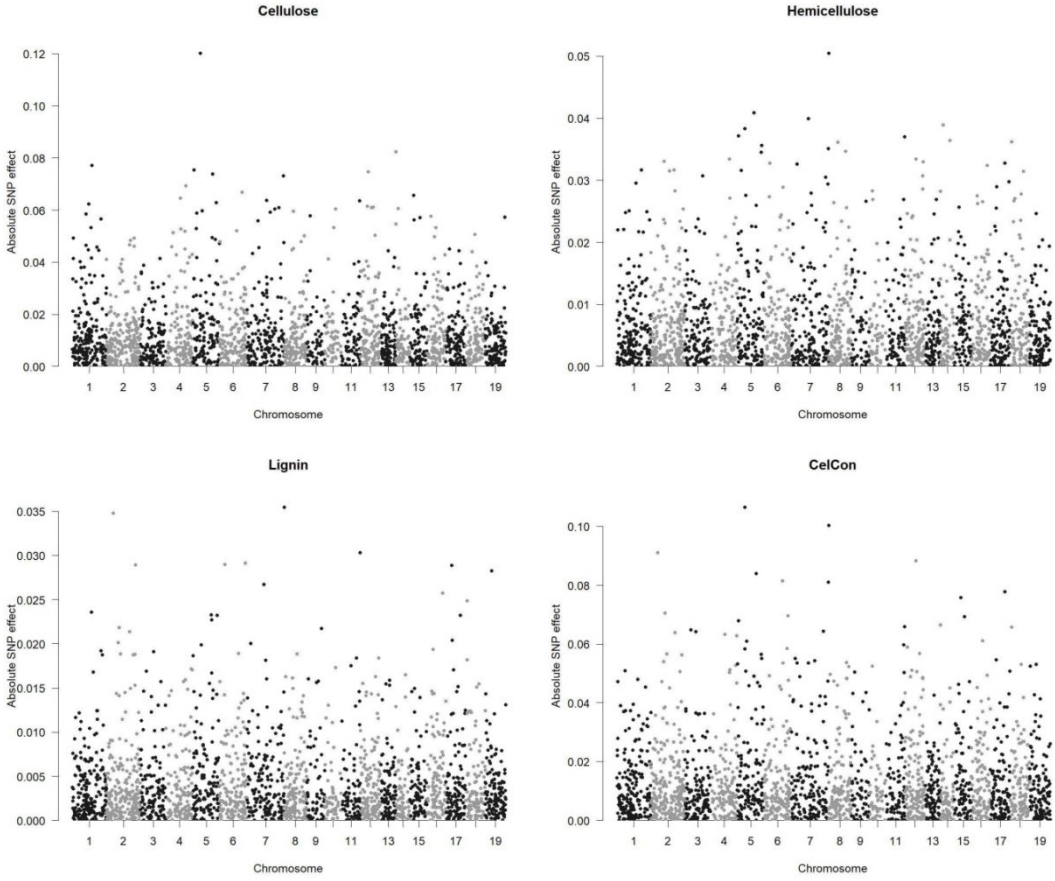


FIGURE 12: MANHATTAN PLOT OF THE ABSOLUTE SNP EFFECTS FOR THE MOST IMPORTANT BIOCHEMICAL TRAITS

In this chapter it has been shown that changing the input of the model and changing parameters will lead to a trait-dependent change in prediction accuracy. For the most realistic prediction accuracies, averages over multiple years should be used, provided that standardized phenotyping protocols are used. High heritabilities have positive effect on prediction accuracies, therefore it should be a goal to obtain deeper knowledge into estimating the most realistic heritabilities. When sufficient high quality markers are available, markers below M.A.F.-threshold should be removed. More markers are expected to give higher prediction accuracies, but a plateau is expected between 10,000-20,000 markers (Slavov et al., 2014). When less genotypes were used in the training population, predictive abilities reduced. Slavov et al. have shown that increasing the training population size will increase prediction accuracies. Therefore it is essential to ensure the size and viability of the experimental population to reach the highest realizable prediction accuracies.

4.4 THE WAY FORWARD

As mentioned before, the end-goal of using genomic prediction is to accelerate *M. sinensis* breeding. The rate of genetic improvement over time consists of four factors and can be calculated using the breeders equation (Eq. 2) (Mackay et al., 2015).

$$R_t = \frac{i*r*\sigma_A}{L} \quad \text{(Equation 2)}$$

where: R_t is the response to selection over time, i is the selection intensity, r is the accuracy of selection, L is the time taken to generate new lines (i.e. generation interval), σ_A is the genetic variance among candidates for selection.

In this study the accuracy of genomic selection has been on average approximately half of phenotypic selection (Table 7). Using genomic selection the generation interval may be significantly reduced, since selection can be performed at an early plant age. If this generation interval is reduced more than half there would already be an increased rate of genetic improvement. Moreover, the amount of selection candidates and its proportion selected (i.e. selection intensity) can be intensified if genomic selection is applied. More crosses and seedlings can be made and the most promising genotypes can be selected an early stage. Other genotypes can be discarded, so there is no need to set-up experimental fields for all individuals.

Within this report several recommendations have been discussed to improve the veracity of the prediction accuracies and other results. Phenotyping and heritability-estimates can be improved by enacting robust phenotyping protocols on a viable population with a randomized design. The most important factors affecting prediction accuracy are the size of the training population and the genetic relationships between training and breeding population (de los Campos et al., 2013; Slavov et al., 2014). These two components should be kept in mind when designing future training populations. It should be noted that actual prediction accuracies can be different due to the leave-one-out validation that was used. When the model is validated using the breeding population (i.e. progeny validation), there will be differences in population structure and a decay in linkage disequilibrium. For future research it is advised to use the phenotypic information from progeny to update the prediction model, since alleles will be fixed by selection and genetic drift. Additionally, it should be considered to extend the training population with data from individuals across multiple environments. This way genomic prediction can help to breed *M. sinensis* for multiple locations and accelerate its utilization as a biobased crop. Actual implementation of all these recommendations requires tremendous effort and will be costly. One component were expenses can be reduced is sequencing costs. After the initial costs of genotyping the training population, the breeding population can be genotyped with a subset of markers. Missing markers will then be imputed from the patterns of recombination and linkage disequilibrium among markers in the training population. In animal breeding, this can be a reduction from 600,000 markers in the training population to as few as 384 in the breeding population (Mackay et al., 2015). Recent suggestions to improve prediction accuracies even further are by usage of indices or incorporation of genomic prediction in crop growth models. For traits that have low heritability and/or predictive ability it is possible to calculate prediction indices based on their genetic correlations to traits with higher heritabilities and prediction accuracies (Mackay et al., 2015). For traits in which genomic prediction was ineffective, the inclusion of even a single trait in selection indices resulted in a several-fold increase in predictive ability (Davey et al., 2017). Since prediction of non-additive gene effects and G*E interactions remains challenging, an integration of crop growth models and genomic prediction (CGM-WGP) was performed (Messina et al., 2017). This resulted in considerably higher environment predictions than traditional genomic prediction and shows potential to predict G*E*M interactions for breeding in future.

5 CONCLUSION

The aims of this study were to characterize the phenotypic variation within the WUR miscanthus collection, obtain insights into the genetic relationships amongst its accessions and to assess the potential of genomic selection. A high degree of phenotypic variation and high estimates of heritability were found for all traits that were measured. This shows the high quality and possibilities the WUR miscanthus collection possesses for breeding novel improved miscanthus accessions. Small imperfections in phenotyping protocols and experimental design were observed. Using robust phenotyping protocols on a viable, randomized experimental population will improve the veracity of the results and ultimately the accuracy of genomic prediction. Sequencing results revealed that accessions with physically close positions had often high genetic relationships. This might lead to an inability to discriminate between genotypic or environmental effects, which can be overcome by implementing a randomized experimental design. RAD-sequencing with a methylation insensitive restriction enzyme resulted in a relative balanced genome coverage, which proposedly has a positive effect on genomic predictions. Alignment of RAD-fragments against the repeat-rich *M. sinensis* genome were found to be challenging and therefore BLAST results are not set in stone. Genomic prediction based on 2600 SNPs resulted in an average prediction accuracy of 0.51. Changing the input of the model and its parameters lead to a trait-dependent change in prediction accuracy. In order to obtain the best and most realistic prediction accuracies, it is advised to use a large set of markers, using accurate heritabilities and average trait-values over multiple years. Recommended future strategies are offspring marker imputation, progeny validation and eventually the integration of various molecular breeding tools. Even better prediction accuracies might be realized by usage of selection indices and/or integration of whole genomic prediction in crop growth models. Considering the long establishment phase of *M. sinensis,* it is expected that implementation of genomic selection will substantially increase the rate of genomic improvement for *M. sinensis*. Although sequencing quality was suboptimal, the findings and pipeline generated in this project will guide future research when high quality sequence data becomes available.

6 REFERENCES

Allison, G.G., Morris, C., Clifton-Brown, J., Lister, S.J. and Donnison, I.S. (2011) Genotypic variation in cell wall composition in a diverse set of 244 accessions of Miscanthus. *Biomass and Bioenergy* **35**, 4740-4747.

Allison, G.G., Robbins, M.P., Carli, J., Clifton-Brown, J.C. and Donnison, I.S. (2010) Designing Biomass Crops with Improved Calorific Content and Attributes for Burning: a UK Perspective. In: *Plant Biotechnology for Sustainable Production of Energy and Co-products* (Mascia, P.N., Scheffran, J. and Widholm, J.M. eds), pp. 25-55. Berlin, Heidelberg: Springer Berlin Heidelberg.

Allwright, M.R. and Taylor, G. (2016) Molecular Breeding for Improved Second Generation Bioenergy Crops. *Trends in Plant Science* **21**, 43-54.

ANKOM_Technology_Corporation (2017a) Method_5_ADF_A200.

ANKOM_Technology_Corporation (2017b) Method_6_NDF_A200.

ANKOM_Technology_Corporation (2017c) Method_8_Lignin_in_beakers.

Arnoult, S. and Brancourt-Hulmel, M. (2015) A Review on Miscanthus Biomass Production and Composition for Bioenergy Use: Genotypic and Environmental Variability and Implications for Breeding. *BioEnergy Research* **8**, 502-526.

Arnoult, S., Mansard, M.-C. and Brancourt-Hulmel, M. (2015) Early Prediction of Miscanthus Biomass Production and Composition Based on the First Six Years of Cultivation. *Crop Science* **55**, 1104-1116.

Atienza, S., Satovic, Z., Petersen, K., Dolstra, O. and Martín, A. (2002) Preliminary genetic linkage map of Miscanthus sinensis with RAPD markers. *Theoretical and Applied Genetics* **105**, 946-952.

Bassi, F.M., Bentley, A.R., Charmet, G., Ortiz, R. and Crossa, J. (2016) Breeding schemes for the implementation of genomic selection in wheat (Triticum spp.). *Plant Science* **242**, 23-36.

Bernardo, R. (2008) Molecular Markers and Selection for Complex Traits in Plants: Learning from the Last 20 Years All rights reserved. No part of this periodical may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. Permission for printing and for reprinting the material contained herein has been obtained by the publisher. *Crop Science* **48**, 1649-1664.

Bhat, J.A., Ali, S., Salgotra, R.K., Mir, Z.A., Dutta, S., Jadon, V., Tyagi, A., Mushtaq, M., Jain, N., Singh, P.K., Singh, G.P. and Prabhu, K.V. (2016) Genomic Selection in the Era of Next Generation Sequencing for Complex Traits in Plant Breeding. *Frontiers in Genetics* **7**, 221.

Bogers, R. (2017) Characterization of the variation of a Miscanthus sinensis collection for biobased end-uses In: *Laboratory of Plant Breeding*. Wageningen: Wageningen University.

Callus, M. (2018) Personal communication - 31-01-2018.

Carpita, N.C. and McCann, M.C. (2008) Maize and sorghum: genetic resources for bioenergy grasses. *Trends in Plant Science* **13**, 415-420.

Clifton-Brown, J.C. and Lewandowski, I. (2002) Screening Miscanthus genotypes in field trials to optimise biomass yield and quality in Southern Germany. *European Journal of Agronomy* **16**, 97-110.

da Costa, R.M.F., Lee, S.J., Allison, G.G., Hazen, S.P., Winters, A. and Bosch, M. (2014) Genotype, development and tissue-derived variation of cell-wall properties in the lignocellulosic energy crop Miscanthus. *Annals of Botany* **114**, 1265-1277.

Davey, C.L., Robson, P., Hawkins, S., Farrar, K., Clifton-Brown, J.C., Donnison, I.S. and Slavov, G.T. (2017) Genetic relationships between spring emergence, canopy phenology, and biomass yield increase the accuracy of genomic prediction in Miscanthus. *Journal of Experimental Botany* **68**, 5093-5102.

de los Campos, G., Hickey, J.M., Pong-Wong, R., Daetwyler, H.D. and Calus, M.P.L. (2013) Whole-Genome Regression and Prediction Methods Applied to Plant and Animal Breeding. *Genetics* **193**, 327-345.

Endelman, J.B. (2011) Ridge regression and other kernels for genomic selection with R package rrBLUP. *The Plant Genome* **4**, 250-255.

Evans, J., McCormick, R.F., Morishige, D., Olson, S.N., Weers, B., Hilley, J., Klein, P., Rooney, W. and Mullet, J. (2013) Extensive Variation in the Density and Distribution of DNA Polymorphism in Sorghum Genomes. *PLOS ONE* **8**, e79192.

Fuss, S., Canadell, J.G., Peters, G.P., Tavoni, M., Andrew, R.M., Ciais, P., Jackson, R.B., Jones, C.D., Kraxner, F., Nakicenovic, N., Le Quere, C., Raupach, M.R., Sharifi, A., Smith, P. and Yamagata, Y. (2014) Betting on negative emissions. **4**, 850-853.

Ghannoum, O., Evans, J. and von Caemmerer, S. (2011) *Chapter 8 Nitrogen and Water Use Efficiency of C4 Plants*.

Gifford, J.M., Chae, W.B., Swaminathan, K., Moose, S.P. and Juvik, J.A. (2015) Mapping the genome of Miscanthus sinensis for QTL associated with biomass productivity. *GCB Bioenergy* **7**, 797-810.

Golfier P, H.F., Zhang W, Voß L, Wolf S, Hell R, Gaquerel E, Rausch T (2016) Lignin from Miscanthus, the Undemanding Giant *24thEUBCE2016* **1DV.1.15 316-324**.

Grabber, J.H. (2005) How Do Lignin Composition, Structure, and Cross-Linking Affect Degradability? A Review of Cell Wall Model Studies This paper was originally presented at the Lignin and Forage Digestibility Symposium, 2003 CSSA Annual Meeting, Denver, CO. *Crop Science* **45**, 820-831.

Heaton, E., Voigt, T. and Long, S.P. (2004) A quantitative review comparing the yields of two candidate C4 perennial biomass crops in relation to nitrogen, temperature and water. *Biomass and Bioenergy* **27**, 21-30.

Heslot, N., Yang, H.-P., Sorrells, M.E. and Jannink, J.-L. (2012) Genomic Selection in Plant Breeding: A Comparison of Models. *Crop Science* **52**, 146-160.

Hodgson, E.M., Lister, S.J., Bridgwater, A.V., Clifton-Brown, J. and Donnison, I.S. (2010) Genotypic and environmentally derived variation in the cell wall composition of Miscanthus in relation to its use as a biomass feedstock. *Biomass and Bioenergy* **34**, 652-660.

Hodkinson, T.R., Klaas, M., Jones, M.B., Prickett, R. and Barth, S. (2015) Miscanthus: a case study for the utilization of natural genetic variation. *Plant Genetic Resources* **13**, 219-237.

Jakob, K., Zhou, F. and Paterson, A.H. (2009) Genetic improvement of C4 grasses as cellulosic biofuel feedstocks. *In Vitro Cellular & Developmental Biology - Plant* **45**, 291-305.

Jannink, J.-L., Lorenz, A.J. and Iwata, H. (2010) Genomic selection in plant breeding: from theory to practice. *Briefings in Functional Genomics* **9**, 166-177.

Jensen, E., Farrar, K., Thomas-Jones, S., Hastings, A., Donnison, I. and Clifton-Brown, J. (2011) Characterization of flowering time diversity in Miscanthus species. *GCB Bioenergy* **3**, 387-400.

Jensen, E., Robson, P., Farrar, K., Thomas Jones, S., Clifton-Brown, J., Payne, R. and Donnison, I. (2017) Towards Miscanthus combustion quality improvement: the role of flowering and senescence. *GCB Bioenergy* **9**, 891-908.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S. and Duran, C. (2012) Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**, 1647-1649.

Lewandowski, I., Clifton-Brown, J., Trindade, L.M., van der Linden, G.C., Schwarz, K.-U., Müller-Sämann, K., Anisimov, A., Chen, C.-L., Dolstra, O., Donnison, I.S., Farrar, K., Fonteyne, S., Harding, G., Hastings, A., Huxley, L.M., Iqbal, Y., Khokhlov, N., Kiesel, A., Lootens, P., Meyer, H., Mos, M., Muylle, H., Nunn, C., Özgüven, M., Roldán-Ruiz, I., Schüle, H., Tarakanov, I., van der Weijde, T., Wagner, M., Xi, Q. and Kalinina, O. (2016) Progress on Optimizing Miscanthus Biomass Production for the European Bioeconomy: Results of the EU FP7 Project OPTIMISC. *Frontiers in Plant Science* **7**.

Liu, S., Clark, L.V., Swaminathan, K., Gifford, J.M., Juvik, J.A. and Sacks, E.J. (2016) High-density genetic map of Miscanthus sinensis reveals inheritance of zebra stripe. *GCB Bioenergy* **8**, 616-630.

Ma, X.-F., Jensen, E., Alexandrov, N., Troukhan, M., Zhang, L., Thomas-Jones, S., Farrar, K., Clifton-Brown, J., Donnison, I., Swaller, T. and Flavell, R. (2012) High Resolution Genetic Mapping by Genome Sequencing Reveals Genome Duplication and Tetraploid Genetic Structure of the Diploid Miscanthus sinensis. *PLOS ONE* **7**, e33821.

Mackay, I., Ober, E. and Hickey, J. (2015) GplusE: beyond genomic selection. *Food and energy security* **4**, 25-35.

McCann, M.C. and Carpita, N.C. (2015) Biomass recalcitrance: a multi-scale, multi-factor, and conversion-specific property. *Journal of Experimental Botany* **66**, 4109-4118.

Messina, C.D., Technow, F., Tang, T., Totir, R.L., Gho, C. and Cooper, M. (2017) Leveraging biological insight and environmental variation to improve phenotypic prediction: Integrating crop growth models (CGM) with whole genome prediction (WGP). *bioRxiv*, 100057.

Meuwissen, T.H.E., Hayes, B.J. and Goddard, M.E. (2001) Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. *Genetics* **157**, 1819-1829.

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., Wicker, T., Bharti, A.K., Chapman, J., Feltus, F.A., Gowik, U., Grigoriev, I.V., Lyons, E., Maher, C.A., Martis, M., Narechania, A., Otillar, R.P., Penning, B.W., Salamov, A.A., Wang, Y., Zhang, L., Carpita, N.C., Freeling, M., Gingle, A.R., Hash, C.T., Keller, B., Klein, P., Kresovich, S., McCann, M.C., Ming, R., Peterson, D.G., Mehboob ur, R., Ware, D., Westhoff, P., Mayer, K.F.X., Messing, J. and Rokhsar, D.S. (2009) The Sorghum bicolor genome and the diversification of grasses. *Nature* **457**, 551.

Phytozome (2018) Miscanthus sinensis v7.1 DOE-JGI,.

Poland, J., Endelman, J., Dawson, J., Rutkoski, J., Wu, S., Manes, Y., Dreisigacker, S., Crossa, J., Sánchez-Villeda, H., Sorrells, M. and Jannink, J.-L. (2012) *Genomic Selection in Wheat Breeding using Genotyping-by-Sequencing*.

Sallam, A.H., Endelman, J.B., Jannink, J.-L. and Smith, K.P. (2015) Assessing Genomic Selection Prediction Accuracy in a Dynamic Barley Breeding Population. *The Plant Genome* **8**.

Schnable, J.C., Freeling, M. and Lyons, E. (2012) Genome-wide analysis of syntenic gene deletion in the grasses. *Genome biology and evolution* **4**, 265-277.

Schnable, P.S., Ware, D., Fulton, R.S., Stein, J.C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T.A., Minx, P., Reily, A.D., Courtney, L., Kruchowski, S.S., Tomlinson, C., Strong, C., Delehaunty, K., Fronick, C., Courtney, B., Rock, S.M., Belter, E., Du, F., Kim, K., Abbott, R.M., Cotton, M., Levy, A., Marchetto, P., Ochoa, K., Jackson, S.M., Gillam, B., Chen, W., Yan, L., Higginbotham, J., Cardenas, M., Waligorski, J., Applebaum, E., Phelps, L., Falcone, J., Kanchi, K., Thane, T., Scimone, A., Thane, N., Henke, J., Wang, T., Ruppert, J., Shah, N., Rotter, K., Hodges, J., Ingenthron, E., Cordes, M., Kohlberg, S., Sgro, J., Delgado, B., Mead, K., Chinwalla, A., Leonard, S., Crouse, K., Collura, K., Kudrna, D., Currie, J., He, R., Angelova, A., Rajasekar, S., Mueller, T., Lomeli, R., Scara, G., Ko, A., Delaney, K., Wissotski, M., Lopez, G., Campos, D., Braidotti, M., Ashley, E., Golser, W., Kim, H., Lee, S., Lin, J., Dujmic, Z., Kim, W., Talag, J., Zuccolo, A., Fan, C., Sebastian, A., Kramer, M., Spiegel, L., Nascimento, L., Zutavern, T., Miller, B., Ambroise, C., Muller, S., Spooner, W., Narechania, A., Ren, L., Wei, S., Kumari, S., Faga, B., Levy, M.J., McMahan, L., Van Buren, P., Vaughn, M.W., Ying, K., Yeh, C.-T., Emrich, S.J., Jia, Y., Kalyanaraman, A., Hsia, A.-P., Barbazuk, W.B., Baucom, R.S., Brutnell, T.P., Carpita, N.C., Chaparro, C., Chia, J.-M., Deragon, J.-M., Estill, J.C., Fu, Y., Jeddeloh, J.A., Han, Y., Lee, H., Li, P., Lisch, D.R., Liu, S., Liu, Z., Nagel, D.H., McCann, M.C., SanMiguel, P., Myers, A.M., Nettleton, D., Nguyen, J., Penning, B.W., Ponnala, L., Schneider, K.L., Schwartz, D.C., Sharma, A., Soderlund, C., Springer, N.M., Sun, Q., Wang, H., Waterman, M., Westerman, R., Wolfgruber, T.K., Yang, L., Yu, Y., Zhang, L., Zhou, S., Zhu, Q., Bennetzen, J.L., Dawe, R.K., Jiang, J., Jiang, N., Presting, G.G., Wessler, S.R., Aluru, S., Martienssen, R.A., Clifton, S.W., McCombie, W.R., Wing, R.A. and Wilson, R.K. (2009) The B73 Maize Genome: Complexity, Diversity, and Dynamics. *Science* **326**, 1112.

Slavov, G.T., Nipper, R., Robson, P., Farrar, K., Allison, G.G., Bosch, M., Clifton-Brown, J.C., Donnison, I.S. and 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**, 1227-1239.

Speight, J.G. (2011) *The Biofuels Handbook*:Royal Society of Chemistry.

Stöcker, M. (2008) Biofuels and Biomass-To-Liquid Fuels in the Biorefinery: Catalytic Conversion of Lignocellulosic Biomass using Porous Materials. *Angewandte Chemie International Edition* **47**, 9200-9211.

Suzuki, M.M. and Bird, A. (2008) DNA methylation landscapes: provocative insights from epigenomics. *Nature Reviews Genetics* **9**, 465.

Swaminathan, K., Chae, W.B., Mitros, T., Varala, K., Xie, L., Barling, A., Glowacka, K., Hall, M., Jezowski, S., Ming, R., Hudson, M., Juvik, J.A., Rokhsar, D.S. and Moose, S.P. (2012) A framework genetic map for Miscanthus sinensis from RNAseq-based markers shows recent tetraploidy. *BMC Genomics* **13**, 142-142.

Taylor, G., Allwright, M.R., Smith, H.K., Polle, A., Wildhagen, H., Hertzberg, M., Bhalerao, R., Keurentjes, J.J.B., Scalabrin, S., Scaglione, D. and Morgante, M. (2016) Bioenergy Trees: Genetic and Genomic Strategies to Improve Yield. In: *Perennial Biomass Crops for a Resource-Constrained World* (Barth, S., Murphy-Bokern, D., Kalinina, O., Taylor, G. and Jones, M. eds), pp. 167-190. Cham: Springer International Publishing.

Trindade, L.M., Dolstra, O., Loo, v.E.N. and Visser, R.G.F. (2010) Plant Breeding and its role in a biobased economy. In: *The biobased economy; biofuels, materials and chemicals in the post-oil era*. Earthscan.

UNFCCC (2015) Adoption of the Paris Agreement. *Report No. FCCC/CP/2015/L.9/Rev.1*.

van der Weijde, R. (2016) Targets and tools for optimizing lignocellulosic biomass quality of miscanthus. Wageningen University.

van der Weijde, T., Dolstra, O., Visser, R.G.F. and Trindade, L.M. (2017a) Stability of Cell Wall Composition and Saccharification Efficiency in Miscanthus across Diverse Environments. *Frontiers in Plant Science* **7**.

van der Weijde, T., Kamei, C.L.A., Severing, E.I., Torres, A.F., Gomez, L.D., Dolstra, O., Maliepaard, C.A., McQueen-Mason, S.J., Visser, R.G.F. and Trindade, L.M. (2017b) Genetic complexity of miscanthus cell wall composition and biomass quality for biofuels. *BMC Genomics* **18**, 406.

van der Weijde, T., Kiesel, A., Iqbal, Y., Muylle, H., Dolstra, O., Visser, R.G.F., Lewandowski, I. and Trindade, L.M. (2017c) Evaluation of Miscanthus sinensis biomass quality as feedstock for conversion into different bioenergy products. *GCB Bioenergy* **9**, 176-190.

VSN_International (2015) GenStat for Windows 18th Edition. p. Web page: GenStat.co.uk. Hemel Hempstead, UK.

Weijde, T.v.d., Alvim Kamei, C.L., Torres, A.F., Vermerris, W., Dolstra, O., Visser, R.G.F. and Trindade, L.M. (2013) The potential of C4 grasses for cellulosic biofuel production. *Frontiers in Plant Science* **4**, 107.

Yu, J., Li, X., Zhu, C., Yeh, C.-T., Wu, W., Takacs, E., Petsch, K., Tian, F., Bai, G. and Buckler, E. (2012) Genic and non-genic contributions to natural variation of quantitative traits in maize. *Genome research*, gr. 140277.140112.

Zhao, H., Li, Q., He, J., Yu, J., Yang, J., Liu, C. and Peng, J. (2014) Genotypic variation of cell wall composition and its conversion efficiency in Miscanthus sinensis, a potential biomass feedstock crop in China. *GCB Bioenergy* **6**, 768-776.

Zhao, X., Zhang, L. and Liu, D. (2012) Biomass recalcitrance. Part I: the chemical compositions and physical structures affecting the enzymatic hydrolysis of lignocellulose. *Biofuels, Bioproducts & Biorefining* **6**, 465-482.

R packages:

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Endelman, J.B. (2011). Ridge regression and other kernels for genomic selection with R package rrBLUP. Plant Genome 4:250-255.

LiLin-Yin (2018). CMplot: Circle Manhattan Plot. R package version 3.3.1. https://CRAN.R-project.org/package=CMplot

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

Turner, S. (2017). qqman: Q-Q and Manhattan Plots for GWAS Data. R package version 0.1.4. https://CRAN.R-project.org/package=qqman

7 APPENDIXES

APPENDIX 1: EXPLANATION OF R SCRIPT USED TO PERFORM GENOMIC PREDICTION

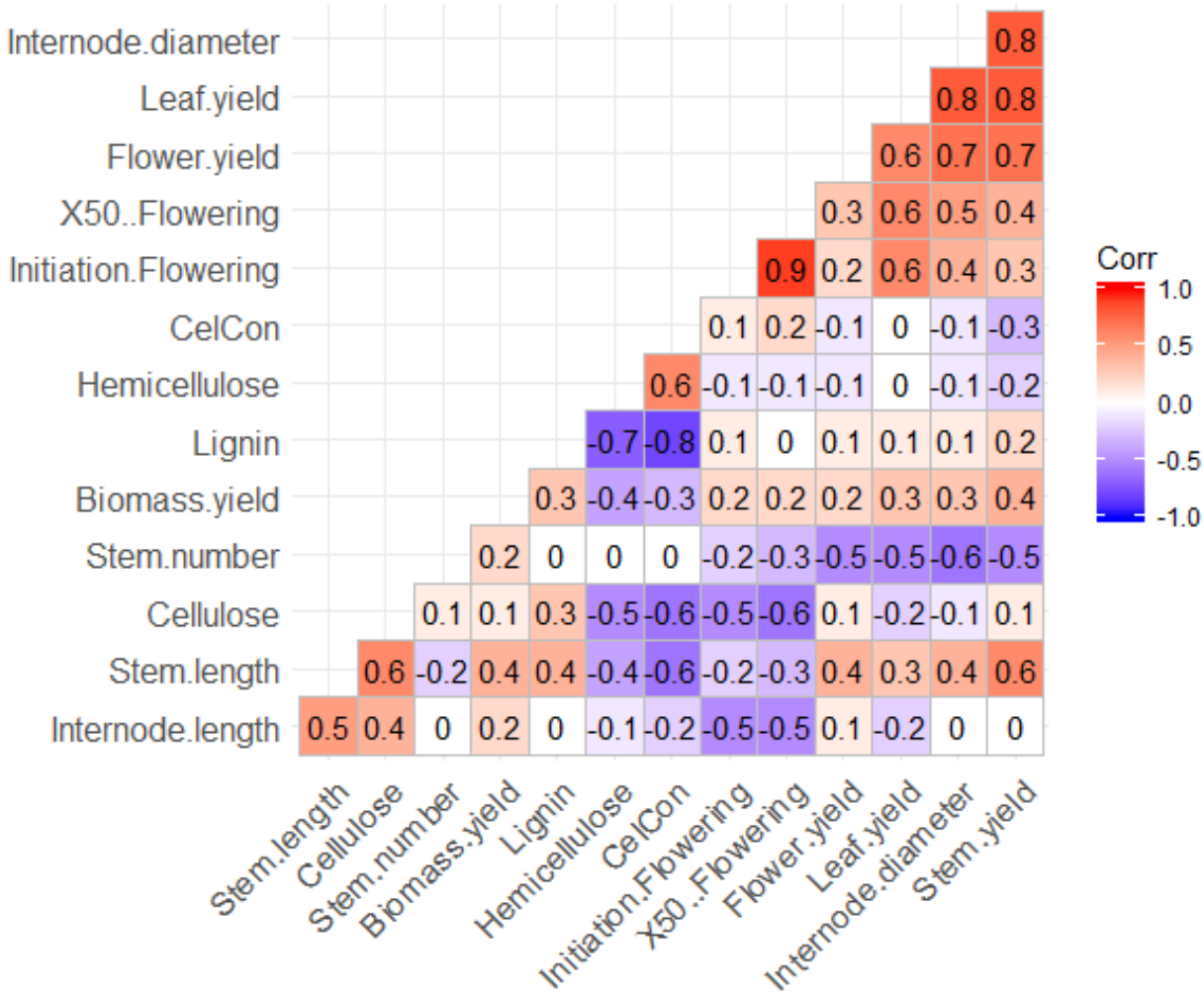
```
# Install and load required packages
install.packages("calibrate")
install.packages("rrBLUP")
library(calibrate)
library(rrBLUP)

# Load marker, phenotype and heritability data from files in your working directory
# (sample files provided via USB)
SNPdata<-read.csv("genotype.csv",header=T,as.is=TRUE,row.names=1)
SNPdata=SNPdata[,-c(1:2)]
SNPdata=SNPdata[,-c(1:2)]
Pheno=as.matrix(read.csv(file = "means.csv",header = TRUE,row.names = 1,as.is = TRUE))
# optional: in our case the genotypes and phenotypes have reverse order, the code below will reverse the
order of the phenotypes
# for(i in 1:ncol(Pheno)) {Pheno[,i] = rev(Pheno[,i])} # for reverse order
# rownames(Pheno) = rev(rownames(Pheno)) # for reverse order
h2list = read.csv("h2.csv")

# Convert SNP data
parseSNPdata = function(x) {
 unique.x = unique(x)
 heterozygote =  setdiff(unique.x,c("A","C","T","G","-"))
 alleles = setdiff(unique.x,union(heterozygote,"-"))
 y = rep(1, length(x))
 y[which(x==alleles[1])] = 0
 y[which(x==alleles[2])] = 2
 y[which(x=="-")] = NA
 return(y)
}
SNPdata_2 = apply(SNPdata,1,parseSNPdata)
rownames(SNPdata_2)=colnames(SNPdata)

# Optional data validation
ifelse(dim(Pheno)[1]==dim(SNPdata_2)[1],"Same number of genotypes in SNPdata and Phenotype
data",stop('Different number of genotypes in SNPdata and Phenotype data'))

# Parameters that are required for the script
nplants = dim(SNPdata_2)[1];nplants
nsnp = dim(SNPdata_2)[2];nsnp
traits = dim(Pheno)[2];traits
p=apply(SNPdata_2,2,mean,na.rm=T)/2
sum_2pq=2*sum(p*(1-p))


####  Make G-matrix
#First scale genotypes to have a mean of 0
SNPdata_3<-scale(SNPdata_2,center=TRUE,scale=FALSE)
#Replace NA's by 0
SNPdata_4<-replace(SNPdata_3,is.na(SNPdata_3),0)
#check positive definiteness of the matrix, and bend if necessary
G<-SNPdata_4%*%(t(SNPdata_4))/sum_2pq
Gt<-eigen(G)
epsilon<-1.e-02
```

```
#matrix Gp can be used to make a heatmap to visualize the relationships
Gp<-G
if(min(Gt$values)<epsilon){
 #  cat(min(Gt$values))
 eig_val<-Gt$values
 eig_val[which(eig_val<epsilon)]<-epsilon
 G<-Gt$vectors%*%diag(eig_val)%*%t(Gt$vectors)
 #  diag(cor(G,Gp))
}

#Compute G-1
G_1<-solve(G)

# Create table to store accuracies that will be generated
accuracy = matrix(nrow=traits, ncol=1,dimnames = list(unlist(dimnames(Pheno)[2]),c("Mean")))

# This code changes the working directory to a subfolder where all generated output will be saved (the
subfolder name is the current time)
wd=getwd()
time= Sys.time()
time = gsub(":",".",time)
wd_tmp = paste(getwd(),"/Output - ",time,"/", sep = "")
dir.create(wd_tmp)
setwd(paste(wd_tmp))


##########
# Below is the code that will generate the output for all plants for each trait.
# It is recommended to select all text  from ### START to ### END (within the R script)
# and press Ctrl+Enter.
##########


### START
# Now compute GEBV
for (r in 1:traits) {
 h2<-h2list[r,2]              #sigma_a^2/(sigma_a^2+sigma_e^2)
 alpha<-(1/h2)-1                    #sigma_e^2/sigma_a^2
 nphen<-nplants                     #all plants have both genotypes and phenotypes

 #set up mixed model equations (MME)
 #1) LHS; dimensions are ((1+nplants)x(1+nplants)); A-1 has dimension (nped x nped)
 LHS<-array(0,c(nplants+1,nplants+1))
 #add 1'1
 LHS[1,1]<-nphen
 #add 1'Z
 LHS[2:(nphen+1),1]<-1
 #add Z'1
 LHS[1,2:(nphen+1)]<-1
 #add Z'Z
 LHS[2:(nphen+1),2:(nphen+1)]<-LHS[2:(nphen+1),2:(nphen+1)]+diag(1,nphen)
 #add G-1*alpha
 LHS[2:(nplants+1),2:(nplants+1)]<-LHS[2:(nplants+1),2:(nplants+1)]+G_1*alpha
 #invert LHS
 LHS_1<-solve(LHS)
```

```
 #2) RHS
 #Read in phenotypes
 y<-Pheno[,r]
 RHS<-array(0,c(nplants+1,1))
 #add 1'y
 RHS[1,1]<-sum(y,na.rm = T) #TEST
 #add Z'y
 RHS[2:(nplants+1),1]<-y
 #3) compute solutions
 SOL<-LHS_1%*%RHS
 GEBV<-SOL[2:(nplants+1)]
 #cor(GEBV,y)


  #4) Now compute solutions without using the phenotype of the individual,
  #       e.g. using leave-1-out crossvalidation (GEBV_CV)
  GEBV_CV<-matrix(0,nplants)
 for (i in 1:nplants){
  #adjust LHS to "remove" plant i
  LHSi<-LHS
  LHSi[1,1]<-LHS[1,1]-1
  LHSi[1,(i+1)]<-0
  LHSi[(i+1),1]<-0
  LHSi[(i+1),(i+1)]<-LHS[(i+1),(i+1)]-1
  #adjust RHS to "remove" plant i
  RHSi<-RHS
  RHSi[1]<-RHS[1]-y[i]
  RHSi[i+1]<-0
  #get new solutions
  SOLi<-solve(LHSi)%*%RHSi
  GEBV_CV[i]<-SOLi[i+1]
 }
 accuracy[r,] = cor(GEBV_CV,y)

 ## For each trait: create a table with the 10 'highest effect' markers and plots of the SNP effects
 SNP_effects<-t(SNPdata_4)%*%G_1%*%GEBV/sum_2pq

dir.create(path =  paste(getwd(),"\\", colnames(Pheno)[r], "\\", sep = ""))
 dir = paste(getwd(),"\\", colnames(Pheno)[r], "\\histogramSNPeffects.jpg", sep = "")
 jpeg(dir)
 hist(SNP_effects,main = colnames(Pheno)[r])
 dev.off()

dir = paste(getwd(),"\\", colnames(Pheno)[r], "\\absoluteSNPeffects.jpg", sep = "")
 jpeg(dir)
 plot(seq(1,length(SNP_effects),by=1),abs(SNP_effects),xlab='SNP',ylab='Absolute SNP-
effect',title(colnames(Pheno)[r]))
 dev.off()

dir = paste(getwd(),"\\", colnames(Pheno)[r], "\\regression.jpg", sep = "")
 jpeg(dir, width = 800, height = 700, res=150)
 plot(GEBV_CV,y,type="n",main = paste(colnames(Pheno)[r]," (R2=",format(round(cor(GEBV_CV,y), 2),
nsmall = 2),")", sep=""),
     xlab="GEBV", ylab="Observed phenotype")
 textxy(GEBV_CV,y,labs = gsub(" OD1302-","",names(y)), offset = 0)
 dev.off()
```

```
 topmarkers = matrix(nrow=10, ncol=2,dimnames = list(c(1:10),c("Marker", "Effect")))
 o = order(abs(SNP_effects), decreasing = T)[1:10]
 topmarkers_o = SNP_effects[o,]
 for (n in 1:10) {
   topmarkers[n,1] = names(topmarkers_o)[n]
   topmarkers[n,2] = round(topmarkers_o[n],8)
 }
 write.csv(topmarkers,file=paste(getwd(),"\\", colnames(Pheno)[r], "\\topmarkers.csv", sep = ""))
}

# Write a table with the mean prediction accuracy for each trait
write.csv(accuracy,file="accuracy.csv")

# Create a heatmap and PCA plot of the genomic relationships
G2=G
diag(G2)=NA
G2[lower.tri(G2)] <- NA
colnames(G2) = rownames(G2) = gsub("X","",colnames(SNPdata))
jpeg("heatmap.jpeg",width=3000,height=3000,res=400)
heatmap(1-G2,Colv=NA,Rowv=NA)
dev.off()

fit <- prcomp(G)
png("PCA.png",units = "px", width = 1466,height = 1066,res = 200)
plot(fit$rotation[, 1:2],type="n",main = "PCA genomic relationships",
    xlab=paste("PC1 (",round(summary(fit)$importance[2,1]%*%100,1),"%)",sep=""),
    ylab=paste("PC2 (",round(summary(fit)$importance[2,2]%*%100,1),"%)",sep=""))
textxy(fit$rotation[,1],fit$rotation[,2],labs=gsub("X","",colnames(SNPdata)))
dev.off()

# Set your working directory back to its original
setwd(paste(wd))
### END
```

APPENDIX 2: ALL PHENOTYPIC CORRELATIONS



APPENDIX 3: FULL PAGE IMAGE OF THE PCA