GENOMIC SELECTION IN RUBBER TREE (*Hevea brasiliensis*) USING SINGLE NUCLEOTIDE POLYMORPHISMS (SNPs) MARKER DATA OBTAINED FROM GENOTYPING BY SEQUENCING (GBS)

By

Norman Munyengwa (R121730B)

A research project submitted in partial fulfilment of the requirements of the Masters of Science Degree in Plant Breeding

Department of Crop Science

Faculty of Agriculture

University of Zimbabwe

June 2019

DECLARATION

FACULTY OF AGRICULTURE

The undersigned certify that they have approved and recommended to the Department of Crop Science for acceptance of the research thesis entitled:

Genomic selection in rubber tree (*Hevea brasiliensis*) using single nucleotide polymorphisms (SNPs) marker data obtained from genotyping by sequencing (GBS).

Approved

Supervisors:

Dr. David Cros

Signature  Date August 21st, 2019

Dr. Edmore Gasura

Signature……...Date……

Department chairperson:

Dr. Edmore Gasura

Signature …….…… Date……....

ABSTRACT

Genomic selection (GS) in rubber tree (*Hevea brasiliensis*) has huge potential to meet future demands of rubber in an economically and environmentally sustainable way. In *Hevea* breeding programmes, genomic selection can be used early in the breeding pipeline to obtain genomic estimated genetic values (GEGVs) for making clonal selections for further large-scale evaluation as potential commercial clonal cultivars. Thus, genomic selection could enhance the efficiency of *Hevea* breeding significantly through decreasing the generation interval and increasing selection intensity, therefore increasing genetic gains per cycle. Within-family genomic selection for rubber latex yield was performed using two sets of 179 and 125 F1 clones from a cross between RRIM600 and PB260 evaluated in two separate phenotypic trials in Côte d'Ivoire. The clones were genotyped using the genotyping-by-sequencing (GBS) approach, which resulted in 3,420 SNPs. A genetic linkage map of the rubber clones was constructed using the JoinMap 5.0 software and two marker imputation methods (Beagle 3.3 and random forest algorithm) were used to impute the missing marker data. The ridge regression best linear unbiased prediction (rrBLUP) was used to predict the GEGVs of clones across-sites. In addition, the effect of marker density on genomic selection accuracy was investigated. Furthermore, the GS accuracies obtained were compared to the GS accuracies obtained using SSR markers and the same phenotypic data. The genetic map contained 1,769 SNPs spanning 2600.9 Centimorgans (cM) and with an average of one SNP in every 1.47 cM. The genetic map also encompassed 308 SSR markers which spanned across 18 linkage groups and with a density of one marker in every 8.4 cM. Beagle imputation performed better than random forest imputation (RFI) as it gave a GS accuracy of 0.52, against 0.48 with RFI. Results also showed that GS accuracy increased with an increase in marker density, and a plateau was reached at 1,000 SNPs with Beagle imputed marker data and at 2,000 SNPs with RFI marker data. The mean between site GS accuracy obtained in this research is similar to the one obtained using SSR markers and the same phenotypic data, opening the way to a cost-effective application of GS in rubber. Results of this study demonstrate that GBS is a rapid, efficient and cost-effective approach for implementing genomics-assisted breeding. This research also showed that GS has high potential to increase yield genetic gain in rubber breeding.

Key words: genomic selection, genomic estimated genetic values, genotyping-by-sequencing, genetic gain, rubber tree.

ACKNOWLEDGEMENTS

Undertaking research work in genomic selection has been a truly life-changing experience for me and it would not have been possible without the tremendous support and guidance from several organizations and people.

Firstly, I would like to express my special gratitude and thanks to my supervisors' Dr David Cros (CIRAD) and Dr Edmore Gasura for guiding and assisting me throughout the whole period of my research study. Without their guidance and constant feedback this research work would not have been achievable.

I would also like to thank Dr Ngalle Hermine Bille and Professor Joseph Martin Bell of the University of Yaounde 1 for always being there for me and making sure that my stay in Cameroon was comfortable. Many thanks also to Dr Catherine Ziyomo who made it possible for me to get a research scholarship from the European Union funded GENES scholarship program. I would like to acknowledge Dr André Clément-Demange (CIRAD), Dr Vincent Le Guen (CIRAD) and Dr Clay Sneller (Oregon State University), as well as my friends Achille Nyouma and Clovis for their motivation throughout the research period.

I gratefully acknowledge the funding received towards my masters' research from the European Union funded GENES project. I thank the *Institut Français du Caoutchouc* (IFC), and the companies Michelin, SIPH, and SOCFIN for their support and for providing the data used for this research, conducted in the framework of the IFC project "Hevea varietal creation in West-Africa". I also thank the companies SOGB (Société Grand Bereby) and SAPH (Société Africaine de Plantations d'Hévéa) for their logistical assistance in the field experiments in Côte d'Ivoire. In addition, I would like to extend my sincere gratitude to the German Academic Exchange Service (DAAD) for funding my masters' studies at the University of Zimbabwe.

I would also like to say a heartfelt thank you to my family for always believing in me and for encouraging me to pursue my dreams.

DEDICATION

This research work is dedicated to my mother Mrs R. Munyengwa. I would not have been where I am today if it was not because of your love and care.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

LIST OF ACRONYMS

CIRAD	Center for International Cooperation in Agronomic Research for Development
DAAD	German Academic Exchange Service
GBS	Genotyping-by-sequencing
GBLUP	Genomic Best Linear Unbiased Estimator
GEBV	Genomic estimated breeding value
GEGV	Genomic estimated genetic value
GS	Genomic selection
GWAS	Genome-wide association studies
IFC	French Rubber Institute
LD	Linkage disequilibrium
MAS	Marker Assisted Selection
M_e	Effective number of independent loci
Ne	Effective population size
NGS	Next Generation Sequencing
QTL	Quantitative Trait Loci
rrBLUP	Ridge regression best linear unbiased prediction
SAPH	Société Africaine de Plantations d'Hévéa
SNPs	Single Nucleotide Polymorphisms
SOGB	Société Grand Bereby
SSRs	Simple Sequence Repeats
SVMs	Support Vector Machines
WGS	Whole-genome sequencing

CHAPTER ONE

1.0 INTRODUCTION

1.1 Background

Rubber tree (*Hevea brasiliensis*, hereafter *Hevea*, 2n = 36), the prime source of natural rubber (*cis*-1,4-polyisoprene), is a preferentially cross pollinating, deciduous perennial crop that belongs to the Euphorbiaceae botanical family (Lau *et al*., 2016). *Hevea* is considered one of the most important crops globally as it provides 99% of natural rubber in the world (Rose and Steinbüchel, 2005), with latex being the major economic product (Pethin *et al*., 2015). Natural rubber production was at 13.28 million tons in 2017 and demand is expected to reach 19.1 million tons by 2025 (Fong *et al*., 2018; Warren-Thomas *et al*., 2015). Although the crop originated from South America in the Amazon rainforest of Brazil, 89% of the world's natural rubber is produced in Asia with Thailand, Indonesia, and Vietnam being the leading producers (Umar *et al*., 2011; Wu *et al*., 2016). Thailand is the leading producer of natural rubber in the world and in 2017 it produced a total of 4.6 million tons (Syahputri *et al*., 2017). After Asia, Africa is the second largest natural rubber producing continent, contributing only 6.7% to world rubber production. Côte d'Ivoire is the 7th highest producer of *Hevea* in the world and the leading producer in Africa. *Hevea* draws its importance on the global market because of its indispensable status as the sole viable source of natural rubber, of which approximately 70% is used in the tire industry with the remaining 30% being used in the manufacture of medical supplies and other industrial purposes (Bandyopadhyay *et al*., 2008).

However, the *Hevea* industry is facing increasing accusations for causing extensive deforestation, accelerating biodiversity loss and increasing carbon emissions, especially in the

top natural rubber producing countries of South-East Asia (Min *et al*., 2017; Warren-Thomas *et al*., 2018). The crop is also facing severe competition for land from other profitable crops especially palm oil, particularly in South-East Asia. Because of the amount of arable land available for cultivation which has become limited due to climate change, desertification and environmental degradation, it has become almost impossible to simply open up untilled land, especially wetlands, to meet rubber production needs (Ronald, 2011). This means that future demands of natural rubber must be met by producing the crop on the same land area as today. There is therefore need for a sustainable solution that increase yields per hectare whilst minimizing environmental impact, and at the same time increasing profits for the poor small-scale farmers who produce approximately 80% of natural rubber globally (Rivano *et al*., 2013). The genetic improvement of natural rubber clonal varieties through genomic selection (Meuwissen *et al*., 2001), which is a modern, state-of-the art approach of marker assisted breeding for quantitative traits, could play a key role to meet this objective.

Genomic selection has emerged as one of the most promising selection strategies to enhance genetic gain, reduce breeding costs and breeding cycle time in tree breeding programs, and its several advantages over both phenotypic and Quantitative Trait Loci (QTL) – based marker assisted selection (MAS) (which depends on marker-trait associations) have been demonstrated (Voss-Fels *et al.,* 2018). Unlike QTL-based marker assisted selection, genomic selection utilizes dense genome-wide markers simultaneously, to predict the genetic values of individuals in the selection population (Wang *et al*., 2018). One key assumption in genomic selection is that a large set of markers are in linkage disequilibrium (LD) with every QTL controlling the phenotypes of interest (Bartholomé *et al*., 2016). Several high through-put and low-cost single nucleotide polymorphism (SNP) chips and next generation sequencing (NGS) technologies such as

genotyping-by-sequencing (GBS) and whole genome sequencing (WGS) platforms have facilitated the production of large amounts of SNP markers for use in genomic selection (Dimitrijevic and Horn, 2018; Li *et al*., 2018).

The first step in genomic selection is to establish a training population which should consist of several hundreds to a few thousand individuals that are related to the validation population and with phenotypes for the traits of interest (Akdemir and Isidro-Sánchez, 2019). The training population is genotyped for a genome-wide panel of markers and also phenotyped for the targeted traits, and a prediction model is developed using these genotypic and phenotypic data. The selection population (for example, 2 weeks old seedlings) is also genotyped but not phenotyped, and the prediction model calculates the genomic estimated breeding values (GEBVs) or genomic estimated genetic values (GEGVs) of the selection population (Edriss *et al*., 2017a). For example, selection among a large number of clones to identify new elite cultivars or clones to be used to advance to the next selection stages could be done at the seedling stage based solely on the GEGVs, thus increasing selection intensity and reducing the generation interval. The main challenge for deriving the genomic selection prediction equation is that the number of SNP effects (*p*) to be evaluated is way higher than the number of individuals (*n*) evaluated (Mei and Wang, 2016). Plant breeders should therefore use an appropriate statistical model that can deal with this problem of 'large *p*, small *n*'. Several statistical models have been developed to calculate GEGVs and to deal with genomic selection challenges (Howard *et al*., 2014). The ridge regression best linear unbiased prediction (rr-BLUP) is among the most commonly used prediction approach. It was proposed by Meuwissen *et al*. (2001) in their ground breaking research paper, and is a parametric model which estimates SNP effects according to a

normal distribution with variance common to all SNPs, thus matching to a genetic determinism following the infinitesimal model.

Besides the statistical model, accuracy of genomic selection can also be affected by the effective population size and marker density, size and structure of the training population, genetic architecture of the traits, relatedness between the training and validation population, level of linkage disequilibrium, trait heritability and method used to impute the missing marker data (as genomic selection models cannot handle them) (Atefi et al., 2016; Rasheed et al., 2017; Wang et al., 2017; Zhang et al., 2019). Genomic data from GBS is characterized by high levels of missing data and several methods such as Beagle and Random Forest (RF) algorithm have been developed to impute the missing data (Chan *et al*., 2016). The Beagle algorithm is widely used to impute marker data. It is a map dependent imputation method which makes use of hidden marker models (HMM) and allele frequencies to impute missing data, thus requiring knowledge regarding marker positions. In rubber tree, due to the scarcity of genomic resources, a physical map corresponding to the 18 chromosomes cannot be achieved currently, and therefore the imputation must be made using a genetic map and a map dependent imputation method such as Beagle version 3.3 (the last version of this software using this type of information). On the other hand, RF is a powerful non-parametric machine learning method which can be used for map-independent imputations, that is, it does not use the linear order of markers to impute missing values.

1.2 Problem Statement

Like all tree breeding programs, the major challenge to rubber breeders is the very long time of the breeding cycle which can last for decades (De Souza *et al*., 2018). The uncertainties associated with planning and conducting such breeding programs can be very high. Conventional

breeding in *Hevea* based on phenotypic selection is less effective for quantitative traits with low heritability, which are influenced by genotype by environment (G×E) interactions (Bhat *et al*., 2016), and is also laborious, require large land and is not cost effective. The three multi-environment and multi-year phenotypic based selection stages (seedling evaluation trials (SET), small-scale clone trials (SSCT) and large-scale clone trials (LSCT)) are the major causes for the high costs and an extended breeding cycle of 20 to 30 years which ultimately limits the number of clonal candidates to be evaluated (Gonçalves *et al*., 2005).

In the SET stage, a large number (around 2000) of full-sib seedlings are evaluated for early identification of traits that are correlated with yield at maturity (Souza *et al*., 2017). The SET stage is the most critical stage for reducing the large number of genotypes issued from hand pollination to a manageable size (less than 200) in the SSCT. However, studies have shown that selection for latex yield at the SET stage using traits such as girth size, latex vessel size, and number of latex vessel rings is less accurate since these traits are poorly correlated with productivity at the adult stage because they are indirect measurements (Gonçalves *et al*., 2005, 2004). Direct quantification of latex productivity is feasible at the SET stage and better correlated with productivity at the adult stage, but not sufficiently enough to predict the potential value of new clones at adult stage. Indeed, the SET is the most inaccurate of all the three selection stages because of the weak juvenile-mature trait correlations and also failure to accurately distinguish between genetic and environmental effects due to G × E interactions.

The use of QTL-based MAS was proposed as an alternative method to shorten breeding cycle time and to enhance genetic gains in tree breeding programs. Although important advances were made in QTL-based MAS, especially for qualitative traits, QTL-based MAS did not make it to the real tree breeding world. The use of QTL-based MAS is ineffective when breeding for

polygenic traits such as latex yield in *Hevea*, and no reliable QTL-based markers have been identified so far (Heslot *et al*., 2015; Robertsen *et al*., 2019; Shamshad and Sharma, 2018). As a result, efforts to apply QTL-based MAS to reduce the generation interval and enhance genetic gains in tree breeding have been fruitless (Crossa *et al*., 2017).

When breeding for quantitative traits, QTL-based MAS has been shown to be inferior even to traditional phenotypic selection (Zhao *et al*., 2014). There is therefore need to explore new and advanced selection methods that are cost effective, fast and with enhanced selection intensity.

1.3 Justification

Genomic selection is a proven technology in both plant and animal breeding to accelerate selection response and genetic gain, but its potential is yet to be fully utilized in perennial crops, and in particular in *Hevea* (Li and Dungey, 2018). Compared to conventional phenotypic selection and QTL-based MAS, genomic selection has high potential to enhance the rate of genetic gain in perennial crops owing to its ability to estimate the genetic values of large numbers of selection candidates early in the breeding pipeline (Sousa *et al*., 2019). In addition, genomic selection allows cheaper and easier selection for late expressing traits and those traits that are difficult to measure such as latex yield, pest and disease resistance, so that more rubber clones can be evaluated than in phenotypic selection, thus allowing for an increase in selection intensity.

Several studies have demonstrated that genomic selection for complex traits is superior to phenotypic selection in terms of genetic gain per cycle and selection response (Beyene *et al*., 2015; Massman *et al*., 2013; Michel *et al*., 2017; Yamamoto *et al*., 2017). Because of the reduced selection cycle time in genomic selection, annual genetic gain is expected to be two to

three times greater than that of conventional phenotypic selection (Sorrells, 2015). For example, Resende *et al*. (2012a) showed that genomic selection can potentially reduce the duration of a conventional eucalyptus breeding program by half, that is, from around 18 to 9 years.

In rubber breeding, the more accurate and efficient genomic preselection could replace the time consuming and costly SETs (Cros *et al*., 2019) and possibly, SSCTs.

However, despite natural rubber's high economic importance and the reported success of genomic selection in other crops, it is surprising that there is only one published article on genomic selection in *Hevea* (Cros *et al*., 2019). The study focused on within-family genomic selection in rubber using simple sequence repeat (SSRs) markers. The article gave very promising results of a selection response increase of 10%. The authors pointed out that the practical implementation of genomic selection in *Hevea* requires a high-throughput and cost-effective genotyping method, which is not the case with SSRs. Therefore, it is the goal of this research to study genomic selection in *Hevea* using GBS marker data from a family of full-sib rubber clones.

1.4 Objectives

1.4.1 General objective

To evaluate the potential of genomic selection in rubber tree clones of a single cross (full-sibs family) using single nucleotide polymorphisms (SNPs) marker data obtained from genotyping-by-sequencing (GBS).

1.4.2 Specific objectives

The specific objectives are:

1. To construct a high-density genetic linkage map of rubber clones of a single family.

2. To compare the performance of two different marker imputation methods (Beagle 3.3 and Random Forest algorithm) on genomic predictions accuracy.
3. To quantify the effect of marker density on genomic predictions accuracy.
4. To compare genomic predictions accuracy obtained with SNP markers and that obtained by using SSR markers and the same phenotypic data.

1.3 Hypotheses

1. Single nucleotide polymorphisms (SNP) marker data from genotyping-by-sequencing can be used to construct a high-density genetic linkage map of rubber clones from a single family.
2. The accuracy of genomic predictions is affected by the method of marker imputation.
3. The accuracy of genomic predictions is affected by SNP density.
4. The accuracy of genomic predictions is affected by the type of markers used.

CHAPTER TWO

2.0 LITERATURE REVIEW

2.1 The rubber tree

Rubber tree (*Hevea brasiliensis* - hereafter referred to as *Hevea*), the prime source of natural rubber, is a deciduous perennial tree of 30 – 40 m high in its natural habitat (the Amazon forest), belonging to the spurge family (Euphorbiaceae) together with other economically important crop species such as the castor oil plant (*Ricinus communis*) and cassava (*Manihot esculenta*) (Souza *et al*., 2018). The leading *Hevea* producing countries in the world are Thailand, Indonesia, Malaysia and China (Tanielian, 2018). Natural rubber is an essential raw material for the tire industry and for more than 50, 000 industrial, health care and household products that have elasticity as a functional attribute (Rahman *et al*., 2013).

Despite the presence of synthetic rubber made from natural gas and petroleum, the importance of natural rubber as a major component of tires used in construction, automotive industry and aircrafts is unmatched due to its unrivalled toughness, resilience, elasticity and resistance to heat build-up (Gonçalves *et al*., 2009). Some of the alternative sources of natural rubber are: Guayule rubber tree (*Parthenium argentatum*), West African rubber tree (*Ficus vogelii*), Ceara rubber (*Manihot glaziovii*), Russian dandelion (*Taraxacum kok-saghyz*), Indian rubber plant (*Ficus elastica*), Panama rubber tree (*Castilla elastica*), False rubber tree (*Holarrhena floribuda*), Lagos silk rubber tree (*Funtumia Africana*), Madagascar rubber tree (*Cryptostegia madagascariensis*), and Palay rubber (*Cryptostegia grandiflora*).

2.1.1 Rubber latex

The latex is harvested by tapping the laticifers (latex vessels), which is a non-destructive latex extraction method that ensures continuous production of the precious colloidal suspension. Tapping is usually done every 2 to 5 days per week for 9 to 11 months each year and the amount of latex obtained from tapping each tree is about 300 ml per year (Sakdapipanich and Rojruthai, 2012). Latex is a colloidal suspension consisting of mainly rubber particles and also proteins, organelles and other non-rubber particles. *Hevea* latex contains two classes of rubber particles: (1) large rubber particles (LRPs) which are surrounded by rubber elongation factors (REF) on their surfaces and account for 94% of rubber particles in latex; and, (2) small rubber particles (SRPs) surrounded by small rubber particle proteins (SRPP) accounting for only 6% of the rubber particles in *Hevea* latex (Berthelot *et al*., 2014). In most cases, latex is harvested from the cups as cup-lumps which are naturally coagulated latex. After the transfer of cup-lumps to the factory where they are stored for a few days or weeks, the cup-lumps are re-processed in the factory in order to wash them (eliminating all impurities), dry them in an air-dryer and are finally pressed to obtain a standardized rubber block with well characterized physical properties. When harvesting is done during the rainy season, formic acid is added to avoid the dilution of latex by the rain in the harvesting cups. It is this processed rubber which is the raw material for manufacturing a countless number of end products with a wide range of industrial, household and health applications.

2.1.2 Plant structure and ecophysiology

The genus *Hevea* consists of 11 inter-crossable species namely *H. brasiliensis*, *H. pauciflora*, H. *guianensis*, *H. camargoana*, *H. nitida*, *H. microphylla*, *H. spruceana*, *H. rigidifolia*, *H. camporum*, *H. paludosa* and *H. benthamiana* (Mantello *et al*., 2012). Mature *Hevea* trees shed

their trifoliate leaves completely for a period of 3 to 4 weeks in a process known as wintering, after which they produce new shoots, leaves and flowers. Flowering normally occurs once a year after leaf shedding and is affected by climatic factors such as rainfall, temperature, latitude and photoperiod (Calle *et al*., 2010). *Hevea* is a monoecious tree, with lateral inflorescences bearing both pistillate and staminate flowers that are greenish to yellow in colour (Yeang, 2007). The rubber tree is considered mature for tapping after six to seven years when it attains a trunk girth of 50 cm at 125 cm height from the ground (Chandrasekhar *et al*., 2005). *Hevea* produces large seeds (3.5 – 6.0 g) that are ovoid in shape with a hard and shiny seed coat that is brown or grey-brown in colour with numerous streaks (Daud *et al*., 2012). The plant has a long juvenile phase of 5 to 8 years before it starts to flower (Dornelas and Rodriguez, 2005).

2.1.3 Harvesting

Rubber latex is harvested from the laticifers by tapping or cutting the bark of the rubber tree using a sharp and specialized knife (Montoro *et al*., 2018). Latex is obtained by periodic tapping every two, three, four or five days depending on the type of clone or labour availability (Liu *et al*., 2016). The most common type of tapping rubber latex is the excision method in which the same cut is reopened at each harvest. With this method, harvesting starts as soon as the trees reach the minimum required girth. Trees grown through budding are harvested when the trees reach a girth of at least 46 cm at a height of 1.5 m from the ground, whilst for trees grown from seedlings latex harvesting starts when a girth of at least 46 cm is reached at a tree height of 75 cm from the ground (as increase in trunk growth is faster in trees grown from seedlings than trees grown from grafting) (Aurélien and Monteuuis, 2017).

Latex production can be enhanced by the application of ethephon once every one or two months, thus increasing the frequency of tapping (Sainoi *et al*., 2017). Tapping creates competition for

photosynthetic assimilates between latex production and growth of the tree resulting in significant reduction in girth growth during the tapping period (Chairungsee *et al*., 2013). Tapping systems around the world are characterized by variations in tapping frequency, concentration and frequency of stimulant (ethephon) application, and tapping cut length.

2.1.4 Rubber breeding objectives

The key breeding objective in *Hevea* is the development of improved high yielding clones with desirable secondary traits such as tolerance to major pests and diseases (*Oidium*, *Colletotrichum*, *Corticium*, *Corynespora*, *Microcyclus*), thick and smooth bark with a good latex vessel system, high growth rate after initiation of latex harvesting, and tolerance to wind (Costa *et al*., 2000). Cultivation of natural rubber has extended into sub-optimal conditions characterized by occurrence of extreme climatic conditions such as prolonged drought, strong winds, low winter and high summer temperatures and also extreme latitudes. Breeding for these abiotic stresses is therefore an important breeding objective in most *Hevea* breeding programmes (Jinagool *et al*., 2015). Such sub-optimum growing conditions also require breeders to develop clones that are high yielding even under high planting densities and poor soil fertility. In Latin America, focus was given on developing clones that are resistant to the South American Leaf Blight caused by the fungus *Microcyclus ulei* (Moraes *et al*., 2012).

Girth or trunk circumference is a measure of vigour and is also considered an economically important trait because it determines the age at which latex harvesting through tapping can begin and is, therefore, important in reducing the uneconomic immature period of the rubber clone (Gonçalves *et al*., 2005). Tapping panel dryness (TPD) syndrome is a physiological disorder which reduces rubber yields and as such developing clones that are tolerant to TPD is also of considerable importance to rubber breeders. Several researches have been done to identify genes

that determine the onset of TPD in rubber (Li *et al*., 2010; Venkatachalam *et al*., 2009). Other breeding objectives depend on socio-economic factors like the availability and cost of labour. In regions where labour is cheap, farmers prefer clones that are adapted to high intensity tapping whilst in areas where labour is expensive, farmers opt for clones that are suited to low intensity tapping.

2.1.5 Rubber breeding and selection

Conventional *Hevea* breeding can be classified into introductions, ortet selection, hybridization and clonal selections. Introductions involve the exchange of rubber breeding material between countries usually under bilateral and multilateral agreements (Ghimiray and Vernooy, 2017). The clones that are already grown at commercial scale in the country of origin are subjected to further evaluations in the receiving country to select those clones that are adaptable to local growing conditions (Jahufer *et al*., 2016). After the agronomic evaluations, promising clones are recommended for commercial planting.

Ortet selection, also known as mother tree selection, is one of the oldest *Hevea* breeding methods (Carron *et al*., 2009). It involves systematic evaluation and selection of high performing and outstanding genotypes that are a result of natural genetic recombination in clonal gardens. Clones developed through ortet selection are known as primary clones (Gonçalves *et al*., 2007). Notable old primary clones that are still grown today include GT1, Tjir1 and PB86.

Hybridization and clonal selection are the most important conventional breeding methods in *Hevea* and have resulted in the release of some outstanding *Hevea* genotypes (Gonçalves *et al*., 2011). With this breeding method, desirable parents are mated to exploit the phenomenon of heterosis and the desirable and selected hybrids are then fixed and maintained easily through

vegetative propagation to produce clonal hybrids. Notable clonal hybrids that are a product of hybridization and subsequent selection are RRIM500 and RRIM600.

Selection in *Hevea* is divided into 3 stages namely seedling evaluation trials (SET), small scale clonal trials (SSCT), and large-scale clonal trials (Costa *et al*., 2000). In the SET, a large number (around 2000) of seedlings issued from hand-pollination are evaluated. Screening at SET is done between and within families and lasts for a period of 2 to 6 years. In the SSCT, few (about 100 to 200) selected full-sib clones from the SET are evaluated in replicated trials for a period of 4 to 8 years. The last stage is the LSCT in which multi-year and multi-locational trials are conducted to evaluate clones on traits such as latex yield, tapping panel dryness, resistance to wind damage and also biotic and abiotic stress tolerance. In the LSCT, only a few promising clones (6 to 20) from the SSCT are evaluated for a period of 15 to 20 years. The three stages can take 20 to 30 years after which an improved clone will be released for commercial plantings (Luke *et al*., 2015).

2.3 Genomic selection

2.3.1 Principle

Genomic selection (GS), initially used in dairy cattle breeding, utilizes whole genome marker data of a phenotyped and genotyped training population to predict the phenotype of a selection population without prior QTL detection as in QTL-based marker assisted selection (Meuwissen *et al*., 2001). The prediction model developed using genotypic and phenotypic data of the training population is used to estimate the genetic values of the selection population for which only genotypic information is available (Resende *et al*., 2012b). The major advantage of genomic selection over QTL-based marker assisted selection is that all markers are incorporated in the prediction model, regardless of the magnitude of their effects, making it one of the most

promising breeding strategies to enhance genetic gain (Cobb *et al*., 2019) for complex traits like latex yield and trunk girth. Genomic selection offers an opportunity for rapid selection of superior genotypes and to shorten breeding cycles and increase selection intensity especially in perennial crops (Cros *et al*., 2015).

The base GS prediction model used to estimate the marker effects is a linear mixed model of the form:

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

Where *y* is the vector of phenotypes of training individuals, *m* is a vector of the random marker effects, β is the vector of fixed effects (for example related to the experimental design, like trials or blocks), Z is the incidence matrix for the vector of marker effects (*m*), i.e. the matrix of genotypes of training individuals, X is the incidence matrix of fixed effects (β), and ε is the vector of residual effects. This model is purely additive (i.e. non additive genetic effects are not taken into account). In this case, the GEBV of the selection candidates ($\hat{g}_C$) are given by:

$$\hat{g}_C = Z_C\hat{m}$$

Where Z_C is the matrix of genotypes of the selection candidates and $\hat{m}$ the vector of marker effects estimated with the data of the training set.

2.3.2 Genotyping-by-sequencing (GBS) in genomic selection

Genotyping-by-sequencing (GBS) follows a modified restriction-site associated DNA sequencing (RAD-Seq; Baird *et al*. 2008) based library preparation protocol for next generation sequencing in an inexpensive and robust multiplexed simple system (Jiang *et al*., 2016). The GBS approach has gained popularity in genomics-assisted breeding as a technology for high

throughput and low-cost genotyping (Wickland *et al*., 2017). Unlike RAD sequencing, GBS library preparation protocol involves fewer steps, require less DNA and it lacks a size selection step (He *et al*., 2014). Important features of GBS include its ability to simultaneously conduct marker discovery and genotyping, and it involves reduced sample handling, no reference sequence limits, fewer PCR and purification steps, efficient barcoding, low cost and easiness to scale-up (Torkamaneh *et al*., 2017). In addition, GBS does not require prior sequencing of the genome and it allows genotyping of plants with complex genomes without prior SNP discovery, thus making the approach more useful for non-model species (Kagale *et al*., 2016) such as *Hevea*. The GBS technique allows for the detection of thousands to millions of SNPs in large plant populations that can be used in linkage mapping, genetic diversity studies, genome-wide association studies and genomics-assisted breeding.

Despite its several advantages, the GBS approach has a few limitations. GBS has a high number of missing values as a result of low depth sequencing (Annicchiarico *et al*., 2015). In addition, GBS is also associated with high levels of sequencing errors (Goonetilleke *et al*., 2017), requiring to make quality controls to remove bad SNPs and possibly bad samples as well. Performing SNP quality control is therefore critical before proceeding to genomic predictions.

In brief, the GBS protocol follows the following steps: normalization of genomic DNA, digestion with restriction enzymes, ligation of barcoded adapter sequences, direct pooling of PCR products (pooling of 96 genotypes to get 1 GBS library), DNA purification on column, PCR amplification with adapter specific primers, double purification of DNA and sequencing on NGS platform (Bhatia *et al*., 2013). The original GBS protocol which utilizes one enzyme ApeKI has been modified in plants into a two-enzyme (Pst1/Msel) GBS protocol to reduce

genome complexity and to enable the development of a uniform library for sequencing (Peterson *et al*., 2014).

2.4 Factors affecting the accuracy of genomic selection

The accuracy of genomic selection is the correlation between the genomic estimated breeding values (GEBVs) or genomic estimated genetic values (GEGVs) and their true values (Lin *et al*., 2014). Prediction accuracy is important in genomic selection due to its linear correlation with genetic gain. Heffner *et al*., (2010) compared the genetic gains per unit time and cost from QTL-based MAS and GS for complex traits in maize and winter wheat using simulated data. Wheat results showed that genetic gain per cycle doubled after increasing GS accuracy from 0.25 to 0.5. Since genomic selection accuracy is correlated with genetic gain per unit of time, it is therefore imperative to explore the various factors that influence the accuracy of genomic selection in plant breeding programs. Factors influencing the accuracy of genomic selection include effective population size and marker density, size and structure of the training population, heritability of traits, genetic architecture, statistical models, relatedness between the training and validation population, linkage disequilibrium, validation approach and method of imputation of missing marker data.

2.4.1 Effective population size and linkage disequilibrium

Effective population size (*Ne*) is the number of randomly mating individuals in a population which lead to the observed rate of inbreeding (Jiménez-Mena *et al*., 2016). A lower effective population size result in higher rates of inbreeding in a population which ultimately leads to genetic drift (Poets *et al*., 2015). One main assumption in genomic selection is that DNA marker coverage is dense enough so that linkage disequilibrium (LD) between the quantitative trait loci (QTLs) and markers will not be broken-up following recombination. The LD between markers

and causal loci is a strong determinant of genomic selection reliability (Meuwissen *et al*., 2001), and *Ne* determines the accuracy of genomic predictions through its effect on LD. In populations with lower *Ne*, LD is high due to higher genetic drift. There is an inverse relationship between LD and the distance between loci. As the distance between loci increases, LD decreases due to more recombination. Generally, strong LD result in higher prediction accuracy (Wientjes *et al*., 2013). Plant and tree breeders can increase the chances of LD between markers and QTLs by deliberately reducing *Ne* (Grattapaglia and Resende, 2011). One way in which genomic selection accuracy can be increased through increasing LD is to use full-sib families, half-sib families and also within family designs (like in the present study).

Schopp *et al*. (2017) assessed genomic selection accuracy within and across bi-parental maize families. The authors trained the GBLUP models with individuals from full-sib, half-sib and unrelated individuals of various training set sizes and varying heritability levels. The authors observed high prediction accuracy within full-sib families (0.41 – 0.97) and for half-sib and unrelated individuals, prediction accuracy was 40 – 60% lower depending with the traits. In addition, Lenz *et al*. (2017) made an assessment of factors that affect genomic selection accuracy for growth and wood quality traits in black spruce (*Picea mariana*). The authors observed a significant reduction in genomic selection model accuracy after using information from half-sibs instead of full-sibs, indicating that the increase in effective population size which was brought about by inclusion of relatedness contributed to higher accuracies. Furthermore, Riedelsheimer *et al*. (2013) investigated on how the training set composition affects prediction accuracy in interconnected bi-parental maize populations. Results showed a 42% decline in genomic prediction accuracy when half-sib double haploid lines replaced full-sib double haploid lines, further indicating the effect of *Ne* on prediction accuracy.

In nature, outcrossing species with various modes of self-incompatibilities such as *Eucalyptus globulus* (McGowen *et al*., 2010) are genetically diverse and have a high *Ne*. Such species have lower LD leading to low prediction accuracy. Beaulieu *et al*. (2014b) observed a sharp decline in prediction accuracy for all traits in white spruce after removing the relatedness between the cross-validation (CV) sets. Researches have showed high genomic selection accuracy when the relationship between the training and validation population was close (Makowsky *et al*., 2011; Meuwissen, 2009).

2.4.2 Marker density and type

Several researches have been conducted to investigate the possibility of using a reduced marker set in genomic selection without reducing prediction accuracies. Plant breeders are interested in the prospects of using a reduced marker set, since it would considerably reduce costs of genotyping for each line in the training set, making it possible to genotype more individuals at the same cost (Robertsen *et al*., 2019). The number of markers (marker density) needed for genomic prediction largely depends on the extent of LD (Werner *et al.*, 2018), which in turn is determined by *Ne* and population structure design. There is a positive linear relationship between marker density and the accuracy of genomic selection. A high marker density usually results in higher prediction accuracy. The reason for an increase in genomic selection accuracy when marker density is increased is because most QTLs will be in LD with some genetic markers and estimates of marker effects will lead to accurate predictions of individual genetic values (Ala Noshahr *et al*., 2018).

Norman *et al*. (2018) investigated the effect and interaction of population structure, training set size and marker density on bread wheat genomic selection accuracy. They showed that genomic selection accuracy was high when they increased marker density, and that high marker density is

more critical when predicting distant relatives. The number of markers required to attain good prediction accuracy is also determined by the relationship between the training population set and the validation set.

Liu *et al*. (2015) published a paper on the effect of genetic relationship and linkage disequilibrium on marker density and size of the training population required to achieve high genomic selection accuracy in maize. Results showed that the closer the real genetic relationship between the training population and the validation population, the fewer the number of markers required to reach a good prediction accuracy. Nielsen *et al*. (2016) reported that a minimum marker set of 1,000 is required to avoid a decline in prediction accuracy. Atefi *et al*. (2016) used simulated animal data to assess the accuracy of genomic selection under various levels of marker density (500, 750 and 1000), trait heritability (0.15, 0.3 and 0.45) and generation intervals of the validation population. Results showed that prediction accuracy increased as the number of markers was increased, reaching the highest when 1000 markers were used. Duangjit *et al*. (2016) reported an increase in prediction accuracy when the number of markers was increased from 500 to 5000 markers. The authors also noted that different traits responded differently to changes in marker density. For example, tomato fruit weight prediction accuracy did not vary much when using different number of markers.

Solberg *et al*. (2008) studied the effects of marker type and density on genomic selection accuracy using simulated data. They compared two marker types (microsatellites and SNPs) and also the use of marker haplotypes in genomic selection versus the use of marker genotypes alone. The different marker densities used were 2; 1; 0.5; and 0.25 N_e markers per morgan using microsatellites and for SNP markers it was 8;4; 2; and 1 *Ne* markers per morgan, where 1 *Ne* markers per morgan meant 100 markers per morgan, if the effective population size (*Ne*) was

100. Results of the study showed that by using microsatellites, accuracy of genomic selection increased from 0.63 to 0.83 after the density was increased from 0.25 *Ne* per morgan to 2 *Ne* per morgan. In addition, by using SNPs, genomic selection accuracy increased from 0.69 to 0.86 after the marker density was raised from 1 *Ne* per morgan to 8 *Ne* per morgan. The authors also noted that it required 2 to 3 times greater density with SNPs to achieve the same accuracy as that of using microsatellites. Using direct marker effects resulted in higher accuracy as compared to using marker haplotypes.

Through assessing the interaction between marker density and population structure on genomic selection accuracy, Norman *et al*. (2018) reported that the response to increased marker density is very high when using a more diverse training population set to predict between poorly related genotypes.

2.4.3 Size and structure of the training population

The size of the training population has been shown to influence the accuracy of genomic prediction in different crops. Nielsen *et al*. (2016) reported a decline in genomic selection accuracy when the number of spring barley lines in the training set was below 200, at which accuracy was more dependent on family structure of the selected training population. Norman *et al*. (2018) studied the effects and interaction of marker density, population structure and population size on genomic selection accuracy in bread wheat (*Triticum aestivum* L.). They used a panel of 10,375 inbred lines genotyped with 18,101 SNP markers. Results showed that prediction accuracy increased with increasing number of training population size, and the increase in prediction accuracy was slow beyond 2,000 lines, indicating that a ceiling will be reached at which an increase in training set size will have little impact on prediction accuracy. The population structure of the bread wheat panel was assessed using K-means clustering and

principal component analysis, and its effect on genomic selection accuracy was assessed using cross-validation analysis according to K-means clusters and breeding cohorts. Results showed that prediction accuracy could be increased if diversity is increased within the training population set, particularly when the relationship between the training and validation population is low.

Sarinelli *et al*. (2019) also reported that an increase in training population size resulted in an increase in genomic selection accuracy in winter wheat. Duangjit *et al*. (2016) evaluated the effect of training population size on genomic selection accuracy in tomato (*Solanum lycopersicum*). Results showed that maximum prediction accuracies were obtained with a training population containing 75% of the tomato accessions. When the size of the training population was reduced from 75% to 25% of accessions, prediction accuracy also decreased.

Zhang *et al*. (2017) reported an increase in genomic selection accuracy when the size of the training population was increased. Cao *et al*. (2017) evaluated the effects of training population size and marker density on genomic prediction accuracy of tar spot complex resistance in maize. The authors noted an increase in prediction accuracy when the training population size was increased in all the maize populations. However, there was a very slow increase in accuracy when the training population was increased from 40% of the population to 90% of the population and the highest accuracy with the smallest standard error was observed when the training population consisted of 60% of the total population, indicating that the optimum training population size was 60% of the total population.

The minimum size of the training population and marker density required to achieve good genomic selection accuracy also depends on the relatedness between the training and validation

population (Gorjanc *et al*., 2017). For example, Bernardo and Yu, (2007b), Lian *et al*. (2014) and Hickey *et al*. (2014) showed that to achieve a prediction accuracy of 0.5 in a bi-parental family, a training set composed of at least 100 phenotyped individuals that have been genotyped at a few hundred markers is needed. These low requirements are because of the low diversity within a family and the high relatedness between the training and validation set found in within-family individuals. However, to achieve the same level of prediction accuracy when dealing with across family individuals, the training set should be composed of a few thousand phenotyped individuals that are genotyped with about 10,000 markers (Hickey *et al*., 2014). The high requirements are because of the high diversity among families and also the low relatedness between the training and validation sets (Pszczola *et al*., 2012).

2.4.4 Heritability of the traits

The higher the heritability of the trait, the greater the accuracy of genomic selection. Atefi *et al*., (2016) reported an increase in genomic selection accuracy from 0.53 to 0.75 when the trait heritability was increased from 0.15 to 0.45. In a study on the implementation of genomic selection in perennial rye grass (*Lolium perenne* L.), Grinberg *et al*. (2016) found out that forage quality traits had the highest prediction accuracy as compared to yield related traits, indicating that the heritability of forage quality traits was higher than that for yield related traits, and hence the higher accuracy. Sarinelli *et al*. (2019) found the same results in winter wheat. Results showed that prediction accuracy amongst all prediction methods were 0.56 for test weight; 0.64 for grain yield; 0.73 for plant height; 0.71 for date to heading and 0.60 for powdery mildew resistance, indicating that accuracy was high in traits with high heritability levels (plant height and date to heading) and low in traits with low heritability levels (test weight and grain yield).

Duangjit *et al*. (2016) also reported high genomic selection accuracy in traits with high heritability levels in tomatoes. Although the research was mostly focusing on the effect of marker density on prediction accuracy, the authors noted that traits that had a greater increase in prediction accuracy with increased marker densities also had higher heritability levels than traits that had low response to increasing marker density. For example, fruit weight (0.814) and soluble solids (0.714) had the highest mean accuracies and the two traits had high heritability values of 0.88 and 0.6, respectively. On the other hand, Aspartate (0.126) and Lysine (0.21) content had the lowest prediction accuracies and were also among the traits with the lowest heritability levels, 0.284 for Aspartate and 0.322 for Lysine.

Zhang *et al*. (2017) evaluated genomic prediction accuracy in 22 bi-parental maize populations. Results showed that an increase in trait heritability, marker density and size of the training population resulted in an increase in genomic prediction accuracy. The authors noted that trait heritability was the most important factor determining prediction accuracy and marker density was the least important.

Viana *et al*. (2016) reported an increase in genomic selection accuracy when trait heritability was increased from 0.3 to 0.7, regardless of marker density and training population size. However, relative accuracy was high in genomic selection as compared to conventional phenotypic selection when the trait heritability was 0.3 and when heritability was raised to 0.7, genomic selection became less efficient than phenotypic selection, indicating that genomic selection is more suitable for traits with low heritability. Heff *et al*. (2011) also noted that the relative accuracy of genomic selection in comparison with conventional phenotypic selection is highest for traits with the lowest heritability.

2.4.5 Statistical models for GS predictions

The major statistical challenge in genomic selection is that the number of markers (*p*) can vastly exceed the number of records (the problem of large-*p* small-*n*) (De los Campos et al., 2013). Different statistical models have been developed for use in genomic selection. Simulation studies have shown clear differences between genomic selection methods in terms of their predictive ability. These methods include the ridge regression (RR) (Whittaker *et al*., 2000), the Bayes A and Bayes B (Meuwissen *et al*., 2001), Bayes C and Bayes Cr (Habier *et al*., 2011) and the Bayesian LASSO (de los Campos *et al*., 2009). These methods differ regarding their assumptions on SNP distribution. The rrBLUP model is a linear parametric genomic selection method which assumes a normal distribution of SNP marker effects with a common variance and a zero mean, and it shrinks all effects equally towards zero using the parameter λ, which controls the trade-off between the model fit and the complexity (Zhao *et al*., 2013). The rr-BLUP is one of the first GS methods proposed by Meuwissen *et al*. (2001), and is one of the most common GS prediction methods. The Bayes A method also assumes that all markers have an effect, but that some of them are in LD with QTLs of moderate to high effects, and therefore have large effects that would not be compatible with the normal distribution. The Bayes B method is similar to Bayes A except that it assumes that some SNPs are in genomic regions where there are no QTLs and thus their effect is zero (Meuwissen *et al*., 2001).

2.4.6 Genetic architecture

It is hypothesized that different prediction methods deal differently with genetic architectures of quantitative traits, i.e. number of QTLs, distribution of their effects and contribution of different gene actions (additive versus non-additive) in traits. Knowledge of the effects of trait genetic architecture and its interaction with statistical models on genomic selection accuracy is important

to plant breeders depending on the breeding objectives. When selecting clones based on *per-se* performance for use in commercial plantings, including models that capture dominance effects is crucial. However, since dominance effects cannot be transmitted to the next generation, breeders should consider models that capture additive effects when their goal is to use clones as parents of new segregating populations (Stich and Van Inghelandt, 2018).

Noshahr *et al*. (2018) studied the effect of different genetic architectures and different genomic selection methods on prediction accuracy. The authors compared three Bayesian methods (Bayes A, Bayes B and Bayesian LASSO) using stochastic simulation across three N_e. Results showed that Bayes B had the greatest selection accuracy for traits influenced by low QTL numbers, low marker density and high N_e. Resende *et al*. (2012a) compared the accuracy of ridge regression best linear unbiased prediction (rr-BLUP), a modified rr-BLUP known as rr-BLUP B, Bayesian A, Bayesian Cπ and Bayesian LASSO in loblolly pine (*Pinus taeda* L.). Seventeen traits with different genetic architectures and heritability, including stem diameter, total height to the base of the live crown, lignin content, wood specific gravity, rooting ability (root number and presence or absence of roots) and *Fusarium* rust resistance were studied. Results showed that Bayes A, Bayes Cπ and the rr-BLUP B had higher prediction accuracy than Bayesian LASSO and rr-BLUP. In addition, the rr-BLUP B model performed equally well as the Bayesian approaches. The performance of rr-BLUP on fusiform rust was expected since the model assumes that all markers contribute equally to observed variation yet the disease is controlled by a few major genes (Sniezko *et al*., 2014). De Almeida Filho *et al*. (2016) observed a significant improvement in prediction accuracy when they shifted from using Bayesian Ridge Regression (BRR) and Bayesian LASSO (BL) to using Bayesian A and Bayesian B in predicting oligogenic traits.

When the contribution of dominance gene action increases, the overall genomic prediction accuracy of prediction models declines (de Almeida Filho *et al*., 2016). This may be because of the inability of prediction models to account for dominance effects the same way they account for additive effects. In addition, epistatic gene action has been shown to be a major contributor of genetic architecture of quantitative traits in model organisms and thus may improve accuracy of predicting breeding values (Morgante *et al*., 2018). Viana *et al*. (2016) reported higher genomic prediction accuracy in a population in which additive variance was 49% higher than the other population. For example, prediction methods incorporating non-additive gene interactions, such as the reproducing kernel Hilbert space (RKHS), are more suitable in predicting how a variety will perform in the future.

2.4.7 Relatedness between the training population and validation population.

The accuracy of genomic selection highly depends on the relatedness between the training and the validation population (Daetwyler *et al*., 2013). A critical parameter in the genomic selection equation is the effective number of independent loci (M_e) (Daetwyler *et al*., 2010). Generally, as the relatedness between the training and validation population increases, prediction accuracy also increases. This is because, the more related the training and validation population is, the lower the effective number of independent loci (*Me*), and the higher the accuracy of genomic selection. Due to its critical role in determining prediction accuracy, several approaches have been proposed for predicting the effective number of independent loci (M_e) and these can be divided into two categories. The first category is the population-based approaches, which are based on the variation of realized relationships (Visscher *et al*., 2006), and which include the two parameters, effective population size (*Ne*) and genome length in morgans (L). The approaches result in expressions for *Me* of: $2N_eL[ln(4N_eL)]^{-1}$, $4N_EL$, and $2N_eL$ where L is the genome

length in morgans (Goddard, 2009; Hayes *et al*., 2009). The total number of independent loci (*Me*) can be calculated by the equation $\frac{2N_eL}{log(4N_eL)}$. The parameter *Me* is very low in full-sibs and thus prediction accuracy is higher than their half-sib counterparts.

Plant breeders can improve accuracy of genomic selection by using training and validation populations that are closely related (Sonah *et al*., 2015). Heffner *et al*. (2011) observed a significant reduction in genomic selection model accuracy after using information from half-sibs instead of full-sibs, indicating the importance of relatedness between the training and validation population.

Ly *et al*. (2013) assessed the effect of relatedness between the training and validation set in Cassava (*Manihot esculenta* Crantz). Two cross-validation schemes were created to evaluate the influence of relatedness on genomic selection accuracy. The first scheme known as cross-validation no close relatives (CV-noCR) was used to avoid closely related clones, whilst cross-validation close relatives (CV-CR) was used to force close relatives between training and validation sets. The relatedness between the training and validation population was measured by identifying for each individual in the validation population, the 10 most closely related individuals in the training population (Clark *et al*., 2012). Results showed that prediction accuracy was the lowest in CV-noCR and the highest was in CV-CR. Edriss *et al*. (2017b) found similar results when they evaluated the effects of population structure, imputation methods and genotype × tester, trial and management interactions on genomic prediction accuracy in a large African maize population. Cross-validation was used to assess prediction accuracy. Prediction accuracy was highest within clusters (0.2 – 0.36) and lowest between clusters (0.04 – 0.26), reemphasizing the effect of relatedness on genomic selection accuracy. Edwards *et al*. (2018)

evaluated the effects of training population design on genomic selection accuracy in wheat. Results showed that using related crosses in training and validation populations resulted in higher prediction accuracies than the use of unrelated crosses, indicating the importance of training population design in genomics assisted breeding.

The effect of relatedness between training and validation population on genomic selection accuracy is also because of the effect of relatedness on LD. The closer the relationship between two populations, the higher the LD (Daetwyler *et al*., 2012). However, the impact of relatedness on prediction accuracy may decrease as the number of SNPs increases.

2.4.8 Validation approaches

Before the practical application of GS, it is necessary to validate the predictions, i.e. to measure the GS accuracy. For this purpose, plant breeders often run cross-validation (CV) schemes within the collected training data (Blonk *et al*., 2010; Meuwissen *et al*., 2001). Cross-validation procedures involve the division of a data set into a training and a validation set, omitting some phenotypes from the prediction model and predicting them using the model.

However, CV can overestimate GS accuracy (Beaulieu *et al*., 2014a; Lorenz *et al*., 2011, p.94; Ly *et al*., 2013b), and therefore validation approaches using independent sites are better (i.e. using one site to train the model and the other site to estimate GS accuracy).

2.5 Marker imputation in Genomic selection

Genotyping by sequencing (GBS) has emerged as the marker platform of choice for genomic selection owing to its high SNP coverage, its low cost and its ability to discover SNPs even for species without a reference genome (Peterson *et al*., 2014; Rasheed *et al*., 2017; Yang *et al*., 2016). Although it can discover thousands to millions of SNPs, GBS is characterized by a high

rate of missing data because of low sequencing coverage which significantly reduces marker density and the number of usable SNPs (Wickland *et al*., 2017). The percentage of missing data depends on library complexity and also depth of sequencing. High depth sequencing results in low proportions of missing data but also in high sequencing costs (Alipour *et al*., 2019). Reducing sequencing depth is cost effective but it comes with high proportions of missing data and leads to a reduction in prediction accuracy (Cericola *et al*., 2018).

Marker imputation is mandatory in genomic selection as the prediction models cannot handle them. Also, it has been shown to be an effective method of mitigating the effects of missing data in genomic selection and genome-wide association studies (Gorjanc *et al*., 2017). Factors that influence imputation accuracy include population structure and the frequencies of marker genotypes in the population (Dassonneville *et al*., 2011). Over the years, a wide variety of imputation methods like Beagle (Browning *et al*., 2018), Impute 2 (Howie *et al*., 2009) and Random Forest algorithms have been developed to account for high levels of missing data in genetic studies. The accuracy of different marker imputation methods may vary under different imputation scenarios due to differences in algorithms and differences in use of information sources. Therefore, it is imperative for plant breeders to select the optimum imputation approach to be used in the population of interest.

2.5.1 Random Forest Imputation (RFI)

Random forests (RF) are considered as one of the most successful and widely used general purpose algorithms used to solve regression and classification problems (Biau and Scornet, 2016). A random forest is a collection of trees, where each of the trees is constructed randomly based on the same tree algorithm (base tree algorithm) and same data set. Each tree in the forest is different due to the inherent randomness of the base tree algorithm. Another source of

randomness in random forest is the process of sub-sampling in which only a sample of the data is used to construct an ensemble of trees. Random forests make predictions by averaging the individual tree predictions in the forest. Random forest is one of the most successful machine learning algorithms owing to its ability to make accurate and robust predictions in a variety of applications (Belgiu and Drăguţ, 2016). When using random forest there are several parameters which need tuning and these include, the number of trees in the forest, choice of base tree algorithm to use, size of leaf nodes, and also the rate of data sub-sampling. Random forest imputation does not require previous information about the order of markers and hence can be implemented without construction of a genetic map (Rutkoski *et al*., 2013).

The RF algorithm starts by sorting markers according to the level of missing data (from lowest to highest). The missing markers are then initialized by sampling the data based on allele frequencies (simple way of imputation), and then a Random Forest regression model is fitted and iterated. One hundred regression trees are grown for each marker vector *y* with missing values using the non-missing values through bootstrapping. In each tree and at each node, a random sample of $\sqrt{n-1}$ predictors is used as splitting variables, in which predictors are other markers at the same row with the missing part of *y*, and *n* is the number of markers. Each tree's terminal node then gives a prediction of the missing part of *y* and the average predictions of the missing part of *y* in all trees are regarded as the imputed values. These steps are repeated until convergence is reached or up to the maximum number of iterations.

2.5.2 Beagle Imputation

Beagle was initially developed for marker imputation in human genetics (Ma *et al*., 2013). Beagle 3.3 is a population-based imputation approach which makes use of linkage disequilibrium (LD) information between the missing SNPs and the observed flanking SNPs to impute missing

data. Beagle is well suited for imputation in unrelated individuals and the approach involves the use of a graphical model to construct a tree of haplotypes present in the training population, and a direct acyclic graph (DAG) is used to summarize the tree by joining its nodes based on haplotype similarity (Sun *et al*., 2012). In Beagle, haplotypes are clustered using the hidden Markov model (HMM). First, Beagle gathers haplotype clusters at each marker and it defines an HMM to get the most likely haplotype pairs based on known genotypes of each individual (Weng *et al*., 2013). Beagle estimates parameters for cluster configuration using empirical frequencies. The next step is estimating the probability of each possible haplotype using genotypic information and the forward-backward algorithm. The last step is a series of iterations, and the default 10 iterations have been shown to obtain high accuracy (Browning and Browning, 2007). The probability of a missing genotype is calculated by simply averaging the posterior genotype probabilities over a series of iterations.

2.6 Experimental results in perennial crops

Cros *et al*. (2019) is so far the only article on genomic selection in rubber tree. The authors studied within family genomic selection for rubber yield using a set of 189 and 143 F1 clones genotyped with 332 simple sequence repeat (SSRs) markers and planted in two separate field trials in Côte d'Ivoire. The effects of statistical genomic prediction methods, size of the training population and marker density on the accuracy of genomic selection was assessed both within and between sites. Between-site genomic selection accuracy was 0.53 when all clones were used in the training population and with all the markers. Results also showed that marker density and training population size strongly affected genomic selection accuracy. In addition, using 300 markers was sufficient enough to achieve a good GS accuracy and increasing the training population size beyond 175 clones would have had a marginal impact on GS accuracy. When

SSR markers with the highest heterozygosity were used, GS accuracy rose to 0.56. Furthermore, genomic selection mathematical models did not affect GS accuracy. Simulation results also showed that implementing genomic pre-selection on 3,000 clones of the considered cross between RRIM600 and PB260 would have raised selection response for rubber latex production by 10.3%. The authors concluded that within-family genomic selection in rubber could lead to the release of more improved rubber varieties which will ultimately lead to higher rubber yields than the current conventional phenotypic based breeding methods.

Cros *et al*. (2015) estimated genomic selection accuracy in oil palm (*Elaeis guineensis*) using SSR markers. Two parental populations (Deli and Group B) involved in conventional reciprocal recurrent selection were used. Each population consisted of 131 individuals and were genotyped with 265 SSRs. Within-population genomic selection accuracies were estimated for the two populations when predicting breeding values of the non-progeny-tested individuals for eight yield traits. Three methods were used for sampling the training sets and the GEBVs were estimated using five statistical methods. Results showed that in Group B, genomic selection could account for both family effects as well as Mendelian sampling terms whereas in Deli it could only account for family effects. Genomic selection accuracy was high ranging from 0.41 to 0.94 and there was a positive correlation between GS accuracy and the relationship between the training set and validation set. The five statistical methods had no effect on genomic selection accuracy. Results also showed that genomic selection can be applied as genomic pre-selection for progeny tests to major yield traits, thus increasing selection intensity.

Cros *et al*. (2017) evaluated genomic pre-selection in commercial oil palm hybrid crosses using SNPs from GBS. The accuracy of GS of seven oil yield components (i.e., annual cumulative bunch production (FFB), annual average bunch weight (ABW), annual cumulative bunch

number (BN), fruit-to-bunch ratio (FB), oil-to-pulp ratio (OP), pulp-to-fruit ratio (PF), and oil extraction rate (OER)) was estimated using A × B hybrid progeny tests with 500 crosses used for training the model and 200 crosses used for independent validation. A panel of more than 5,000 SNPs from GBS was used for genomic preselection. The GBLUP was used to perform GS predictions using SNP data of both the training and validation population and phenotypic data of the training crosses. Results showed that prediction accuracies increased with an increase in marker density up to 500 SNPs. Prediction accuracies started to plateau from 500 SNPs up to 2,000 SNPs. Prediction accuracies varied from high (0.73) to low (0.28) depending on traits. GS was able to capture genetic differences that were present within families, and it required at least 2,000 SNPs with less than 5% missing data, and imputed using pedigree information. The authors concluded that genomic preselection could have had increased the selected hybrids bunch production yield by more than 10%.

Kwong *et al*. (2017) evaluated the effect of two marker systems and eight modelling methods for implementing genomic selection in Nigerian dura × Deli dura family with 112 individuals. The traits selected were shell-to-fruit (S/F), mesocarp-to-fruit (M/F), fruit-to-bunch (F/B), kernel-to-fruit (K/F), oil per palm (O/P) and oil-to-dry mesocarp (O/DM). The two marker systems evaluated were single nucleotide polymorphisms (SNPs) and simple sequence repeats (SSRs). Eight genomic selection modelling methods used for estimating the accuracy of genomic selection for the traits were Ridge Regression, RR-BLUP, Bayesian A, B, Cπ, LASSO and two machine learning methods (Random Forest and Support Vector Machine (SVM)). Results showed that O/DM had the highest genomic heritability whilst O/P and P/B had the lowest. Genomic selection accuracies were low with SSRs, with trait accuracies around 0.20. The average genomic selection accuracy of the two machine learning methods was relatively higher

(0.24), as compared to 0.20 achieved by other GS methods. Traits with the lowest mean accuracies were M/F and O/P (0.18), whilst F/B (0.28) had the highest accuracy. The accuracies for all traits were improved by using genome-wide SNPs especially for M/F (0.30), S/F (0.39) and O/DM (0.43). The average genomic selection prediction accuracy of the two machine learning methods was 0.32, as compared to 0.31 of the other methods.

Ratcliffe *et al*. (2015) studied the potential of genomic selection in interior spruce (*Picea engelmannii* × *glauca*) breeding utilizing a genotyped population of 769 spruce trees derived from 25 open-pollinated families. Repeated tree height measurements were done at ages 3, 6, 10, 15, 30, and 40 years to allow the temporal testing of genomic selection methods. Single nucleotide polymorphism (SNP) discovery was done using the genotyping-by-sequencing (GBS) pipeline for non-model species. Three unordered marker imputation methods (K-Nearest Neighbor with special family weighting (KNN), mean imputation (M60) and singular value decomposition (SVD)) were used to impute the data which had 60% missing information. Three genomic selection models were evaluated based on their predictive accuracy, and their subsequent marker effects. Prediction accuracy was moderate (0.31 to 0.55) and were of enough capacity to deliver enhanced selection response over traditional pedigree-based selection. Temporal genomic selection prediction accuracy decreased with increasing difference in age between the training population and validation set (0.04 – 0.47). Imputation results showed that SVD and KNN yielded a higher number of SNPs and had higher prediction accuracies than mean imputation (M60). In addition, the Bayes Cπ and ridge regression BLUP (rrBLUP) yielded the same level of prediction accuracy and performed better than the generalized ridge regression heteroscedastic effect model for the traits under study.

Muranty *et al*. (2015) studied the accuracy and responses of genomic selection (GS) for key traits in apples. Prediction accuracy and selection response was assessed for key culling traits namely: fruit cropping, pre-harvest dropping, per cent of russet, attractiveness, fruit size, and four components of skin colour, over-colour, per cent over-colour, ground colour and type of colour. The training population consisted of 977 individuals derived from 20 pedigreed full-sib families. Historic phenotypic data for 10 traits related to fruit productivity and external appearance were available. The genotypic data for the 7,829 SNPs was generated with an Illumina 20K SNP array. A genome-wide prediction model was built using these data and was used to calculate the genomic breeding values of five application full-sib families. These five application families were phenotyped for one year and their phenotypic values were compared with the predicted breeding values. Results showed that genomic prediction accuracy for the 10 traits reached a maximum value of 0.5 with a median value of 0.19. In addition, GS accuracies were strongly affected by heritability of traits and phenotypic distribution. Significant selection response was observed for traits with symmetric phenotypic distribution and high heritability. Non-significant response was observed on traits with low heritability or traits with reduced or skewed phenotypic variation. Furthermore, the degree of relatedness between the training and validation population did not affect prediction accuracies among the five application families. They concluded that genomic prediction has huge potential to accelerate breeding progress in fruit tree crops by overcoming the long generation intervals and high phenotyping costs.

Iwata *et al*. (2013) evaluated the prospect of applying genome-wide association studies and genomic selection in Japanese pear (*Pyrus pyrifolia*) breeding. The study used 76 pear cultivars to detect significant associations of 162 DNA markers with nine agronomic traits which include harvest time, fruit size (fruit weight), resistance to black spot, fruit shape in longitudinal section,

sugar content, acid content, vigour of tree and number of spurs. Multi-locus Bayesian regression models accounting for categorical phenotypes were applied for both GWAS and GS model training. Significant marker-trait associations were observed at harvest time, number of spurs and black spot resistance and two associations were closely linked to known loci. Whole genome predictions for genomic selection were highly accurate (0.75) for harvest time, at moderate levels (0.38 – 0.61) for fruit shape in longitudinal section, resistance to black spot, fruit size, number of spurs, firmness of flesh and acid content and were low (< 0.2) for sugar content and tree vigour.

Beaulieu *et al*. (2014) assessed genomic selection accuracies between and within environments and breeding groups (BG) in white spruce. Genomic prediction accuracies for growth and yield traits were determined using 1748 trees and 6,932 single nucleotide polymorphisms (SNPs). Each of the breeding groups had an effective size of Ne ~ 20 and marker subsets were also tested. Results showed that cross-validation (CV) prediction accuracies for least absolute shrinkage and selection operator and the ridge regression (RR) models reached those of pedigree-based models. In addition, with strong relatedness between CV sets, prediction accuracies for RR within environment and breeding group were high for wood ($r = 0.71 – 0.79$) and moderately high for growth traits ($r = 0.52 – 0.69$) depending on the heritability of the traits. Accuracies for both classes of traits reached between 83% and 92% of those obtained with phenotypic and pedigree data. Prediction accuracy in untested environments was moderately high for wood ($r \geq 0.61$) and dropped significantly for growth traits ($r \geq 0.24$), pointing out the need for phenotyping in all test environments and to model genotype-by-environment interactions (G×E) for growth traits. In addition, prediction accuracies for all traits and sub-populations decreased sharply, as the relatedness between CV sets was removed. They concluded

that in order to obtain good prediction accuracies, high relatedness between CV sets is needed, and genomic selection models should be built within the same breeding population only.

Li and Dungey. (2018) evaluated the potential gains of implementing genomic selection over forward selection in conifer breeding using stochastic simulation. The authors selected several methods to speed up deployment of selected material and these include using additional replicates of conifer clones in archives for crossing, top-grafting on mature seed orchard ortets, embryogenesis and clonal propagation. Results showed that genetic gain per generation increased when the training population size was large (800 c.f. 3000 clones) and/or when the heritability of the traits was higher (0.2 c.f. 0.5). The largest genetic gain of 24% was realized when a large training population size (3000 clones) and high traits heritability (0.5) were combined. In addition, the accuracy of GEBVs increased with an increase in training population size, heritability of the traits, and also SNP marker density. Furthermore, results of calculated prediction accuracies of genetic gain per unit time and GEBVs suggested that a minimum training population size of 2000 clones is required for effective genomic selection of conifers. They concluded that with genomic selection with a training population size of 2000 clones and a 60K SNP panel, breeders can expect a 1.58 mm per year increase in diameter-at-breast-height (DBH) and 2.44 kg/m^3 per year increase for wood. The authors concluded that deploying genetic material (clones) selected using genomic selection with top-grafting for early cloning could be the best option in forest tree breeding.

Suontama *et al*. (2019b) studied the potential of genomic selection across two Eucalyptus nitens breeding populations with varying selection histories. A breeding population consisting of 691 individuals representing two seed orchards with varying selection histories were genotyped using a high-density SNP chip (EUChip60K). Records of phenotypic data on growth and form traits,

and for wood quality traits at age seven were available. The GBLUP was used to build the prediction model, which was compared to the traditional pedigree-based alternative using the ordinary BLUP. Results showed that substantial improvement of genetic gain and breeding value accuracy can be achieved with GBLUP. In addition, cross-validation within and across two different seed orchards showed that higher GS predictive accuracy can be achieved through increasing the training population size.

CHAPTER THREE

3.0 MATERIALS AND METHODS

3.1 General overview

The study used phenotypic and genotypic data of 304 *Hevea* clones from the F1 cross between PB260 × RRIM600. The clones were evaluated in two different small-scale clonal trials (SSCTs) in Côte d'Ivoire, with 179 clones at HR46 and 125 at Sapest13 under the Center for International Cooperation in Agronomic Research for Development (CIRAD) and French Rubber Institute (IFC) rubber breeding programme. The two SSCTs were implemented using conventional experimental designs, which enabled the reliable estimation of clone values (phenotypes). In addition, the clones were genotyped by GBS, resulting in 3,420 SNP markers. Genomic selection (GS) models trained using genome-wide marker data and phenotypes of clones in one site were used to predict the phenotypes of clones in another site (across-site predictions). Across-site predictions were performed to test the potential of replacing phenotyping by GS predictions. A

high-density genetic linkage map was constructed prior to Beagle imputation and the effect of two marker imputation methods (Beagle 3.3 and Random Forest Imputation) on GS accuracy was quantified. The effect of marker density on GS accuracy was investigated.

3.2 Study sites

The field experiments for phenotyping were conducted in the south-western parts of Côte d'Ivoire. Experiments comprised two study sites, namely Sapest13 located at Société Africaine de plantations d'Hévéas (SAPH) estate and HR46 located at Société des Caoutchoucs de Grand-Béréby (SOGB) estate. HR46 is characterized by gravelly clay-loam soils and it lies at an altitude of 33 meters above sea level (m.a.s.l) with a longitude of 7° 06' 05" W and a latitude of 4° 40' 54" N. Sapest13 has deep sandy soils and it has a longitude of 4° 36' 39.74" W, a latitude of 5° 19' 47.79" N and it lies at an altitude of 89 m.a.s.l. Both sites experience the same tropical climatic conditions, with an annual average temperature of 26°C and annual average rainfall of 1,600 mm.

3.1.2 Planting material and field phenotyping

A total of 304 F1 clones were phenotyped in the study, with HR46 and Sapest13 having 179 and 125 clones, respectively. There were two common clones in both sites, thus making a total of 304 clones altogether. Data from the two common clones was only used to train the genomic prediction model and excluded from the validation sets.

The 304 F1 clones belong to a cross between the secondary rubber clone RRIM600 and PB260, in which RRIM600 was used as a male and PB260 as a female. RRIM600, a widely grown and universally adapted accession from Malaysia which originated from a cross between primary rubber clones PB86 and TJI1, is a high yielding clone even under sub-optimal environments.

PB260, also a Malaysian clone, was derived from a PB5/51 (PB56 × PB24) × PB49 cross, and is a high yielding clone which is highly fertile and also considered suitable for sub-optimal areas. The planted ramets were produced under nursery conditions by grafting on root stocks generated from seeds produced from natural pollination of primary clone GT1 of Indonesia. The average number of ramets per clone at HR46 and Sapest13 was 11 and 13, respectively, giving a total of 2,016 and 1,869 ramets per site, respectively.

The HR46 trial was planted in July 2012 and the Sapest13 trial was planted in July 2013. Ramets were planted at a spacing of 2.5 × 2.5 meters with a plant population of 1,600 per hectare. The research used an almost complete block design with individual trees randomized within each of the 6 blocks. The trait under study was rubber latex production.

Data collection on rubber yield was recorded in both sites and for each ramet by tapping 3 times a week for 3 consecutive months. In HR46, tapping started at the end of the dry season at 32 weeks after planting whilst at Sapest13 tapping started 38 weeks after planting, at the end of the rainy season. At HR46, each tree had a mean cumulative partly-dried rubber yield of 78.7 g, which ranged from 0.50 to 318 g per tree, meanwhile at Sapest13 the mean per tree was 244.6 g, with a range of 0.25 to 840.1 g per tree. To get the clone values using raw ramets data from the two sites, a linear mixed model and Best Linear Unbiased Predictor (BLUP) analysis was carried out using ASReml-R version 3.0 statistical package (Butler *et al*., 2009). The clone values were in adjusted forms so as to cater for effects related to blocking and also variations in size among the trees at the time of tapping. In this research study, the adjusted clone values are referred to as phenotypes. The broad sense heritability (H^2) at each site was 0.9, and was calculated as per equation below:

$$H^2 = \sigma_G^2 / \left(\sigma_G^2 + \frac{\sigma_E^2}{h_r}\right)$$

Where σ_G^2 is the genetic variance of clones, σ_E^2 is the residual error variance and h_r is the trial harmonic mean number of ramets per clone (Gonçalves et al., 2006), with σ_E^2 and σ_G^2 obtained from the linear mixed model.

3.2 Marker genotyping

3.2.1 Genomic DNA extraction

Young and healthy leaflets were collected and genomic DNA was extracted from leaf slices of 6 mm in diameter, each ring weighing 1 mg. Extraction of high-quality genomic DNA was done on Macherey-Nagel magnetic beads, using the Beckman robot. In brief, the process of DNA extraction was done in three stages as follows:

Grinding of leaf samples: the leaf samples were placed in a 96-well Corning deepwell of 1.1 ml round wells in the presence of a 3 to 4 mm diameter ball. The samples were then frozen and crushed in liquid nitrogen using a Genogrinder ball mill.

Lysis stage: the buffer was preheated to 72 ° C and the leaf sample powder was mixed with 400 µl of hot extraction buffer. The mixture of leaf sample powder and hot extraction buffer was shaken by vortexing and incubated at 72 ° C for one hour in an oven. This was followed by 30 minutes of high-speed centrifugation at 4000 rpm using an Eppendorf 5810R centrifuge.

Purification: this process was performed on the Beckman robot which features fully automated milling, leaching and eluting steps. The automated purification protocol proceeded as follows:

1. Preparation of binding on magnetic beads in a deepwell by mixing 20 µl of diluted magnetic beads and 300 µl of isopropanol.

2. Removal of 300 µl of supernatant and addition to the binding buffer followed by incubation.
3. Ringing of magnetic beads on the magnetic support and elimination of the supernatant.
4. Washing cycles using wash buffers (MC3, MC4) and ethanol (EtOH80) followed by addition of the elution buffer (MC6).
5. Use of the program 02_elution and recovery of the supernatant in PCR96 Sorrenson plates.

3.2.2 Genotyping-By-Sequencing (GBS)

After DNA extraction, the DNA was digested by two enzymes Pst1 and MSE1, in league with the barcode and adapters, and multiplexed to 96 individuals per bank. The banks were sent to GENEWIZ (USA) for sequencing using a next generation sequencing technology known as Genotyping-By-Sequencing (GBS) (Elshire *et al*., 2011). In brief, the GBS protocol followed the following steps: normalization of genomic DNA, digestion with restriction enzymes, ligation of barcoded adapter sequences, direct pooling of PCR products (pooling of 96 genotypes to get 1 GBS library), DNA purification on column, PCR amplification with adapter specific primers, double purification of DNA and sequencing on NGS platform (Bhatia *et al*., 2013).

After sequencing, the SNP data was sent back to CIRAD in the form of Fastq files that were then transformed to the Variant Call Format (VCF) file. Indels and SNPs that were not biallelic were removed from the VCF file using VCFtools (Danecek *et al*., 2011). All SNPs with a minor allele frequency of less than 15% were also removed, as this was not compatible with the possible segregation patterns expected with biallelic markers in a single cross. In addition, SNP datapoints with a read depth less than eight were set as missing and SNPs with more than 50%

missing data were removed. This resulted in a raw VCF file with 83,259 SNPs, which was provided by CIRAD to conduct the present thesis work.

3.3 Production of the final SNP dataset

To get the final VCF file from the VCF file converted from the Fastq format in order to perform genomic predictions, the following steps were followed to ensure suitable molecular data quality:

1. Removal of 30 illegitimate individuals (identified in a separate study) from the VCF file using VCFtools in Linux.
2. Removal of 396 SNPs that were monomorphic. This resulted in a VCF file with 82,863 SNPs.
3. Thinning of SNPs to keep only one SNP per window of 500 base pairs apart. This resulted in a VCF file with 35,802 SNPs.
4. A histogram of percent missing data per individual was plotted in R and the five individuals with percent missing data greater than 50% were removed from the VCF.
5. A histogram of mean read depth (DP) per SNP was plotted in R, and 116 SNPs that appeared as outliers, that is, with a mean DP greater than 400 were identified and removed from the VCF file, as it was assumed that they were found in duplicated regions of the genome. This resulted in a VCF file with 35,686 SNPs.
6. Separating the VCF file into parental VCF and progeny VCF.
7. Each of the two parents was replicated three times in the VCF file and a script was run in R to determine the true genotypes of the two parents. The genotype which appeared at least twice amongst the 3 replicates of the parents was chosen as the true genotype and the other genotype(s) (if any) were regarded as genotyping errors. In the case in which all the three loci were different, the data point was set as missing data.

8. The percentage of heterozygous genotypes per SNP was computed in R using the following formulae:

$$\frac{Number\ of\ heterozygotes\ per\ SNP}{Number\ of\ samples - Number\ of\ missing\ data}$$

   A histogram of percentage heterozygosity was plotted and all SNPs with heterozygosity percentage greater than 80% and less than 20% were removed both in the parental VCF and the progeny VCF, as this was not compatible with the possible values expected with biallelic markers in a single cross. This resulted in parental and progeny VCF files with 33,517 SNPs each.
9. Removal of SNPs that did not pass the comparison test of expected and observed segregation (0/1 × 0/1, 0/1 × 1/1, 0/1 × 0/0, 1/1 × 0/1, 0/0 × 0/1) using a p value < 0.01, following the Monte Carlo exact multinomial test. This was done with the function '*multinomial.test'* in the EMT R package (Lawal, 2003) in both the parental and progeny VCF files, and resulted in 3,458 SNPs.
10. Removal of SNPs that were homozygous in the two parents with different alleles (AA × BB) as they were not compatible with the possible segregation patterns expected with biallelic markers in a single cross

The histograms that were plotted in R to show the distribution of percentage missing data per SNP, percentage of missing data per individual, mean read depth per SNP, and percentage heterozygosity after the above quality assurance steps are shown in Figure 3.1 below.

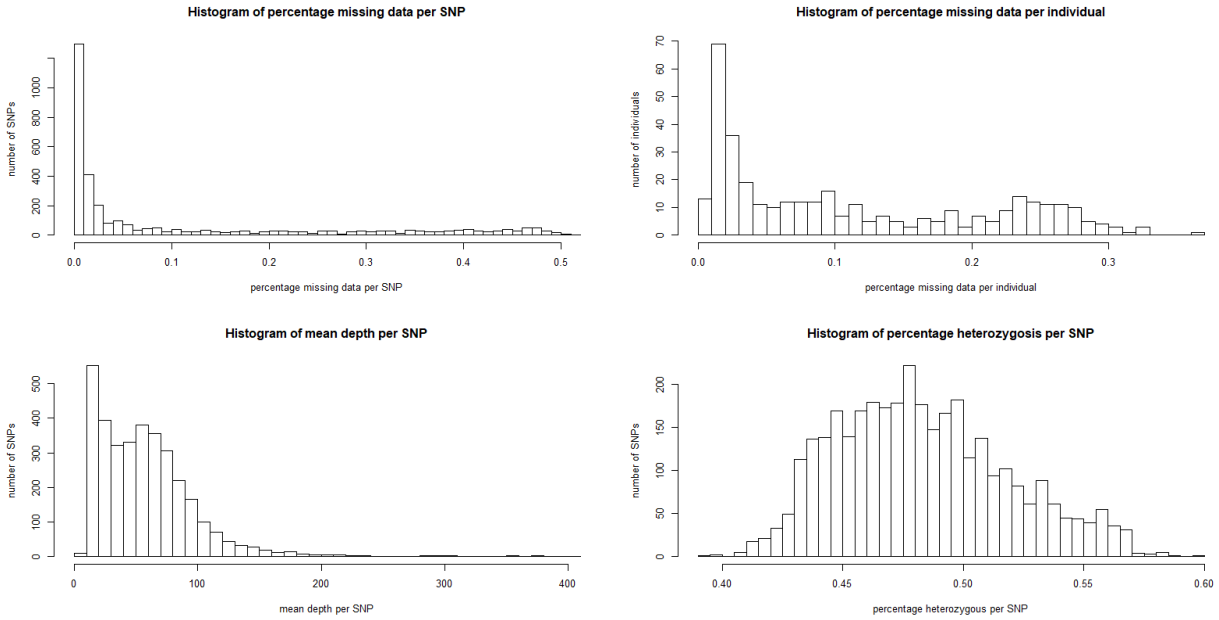


Figure 3. 1: Distribution of percentage missing data per SNP (top left), percentage missing data per individual (top right), mean depth per SNP (bottom left), and percentage of heterozygous genotypes per SNP.

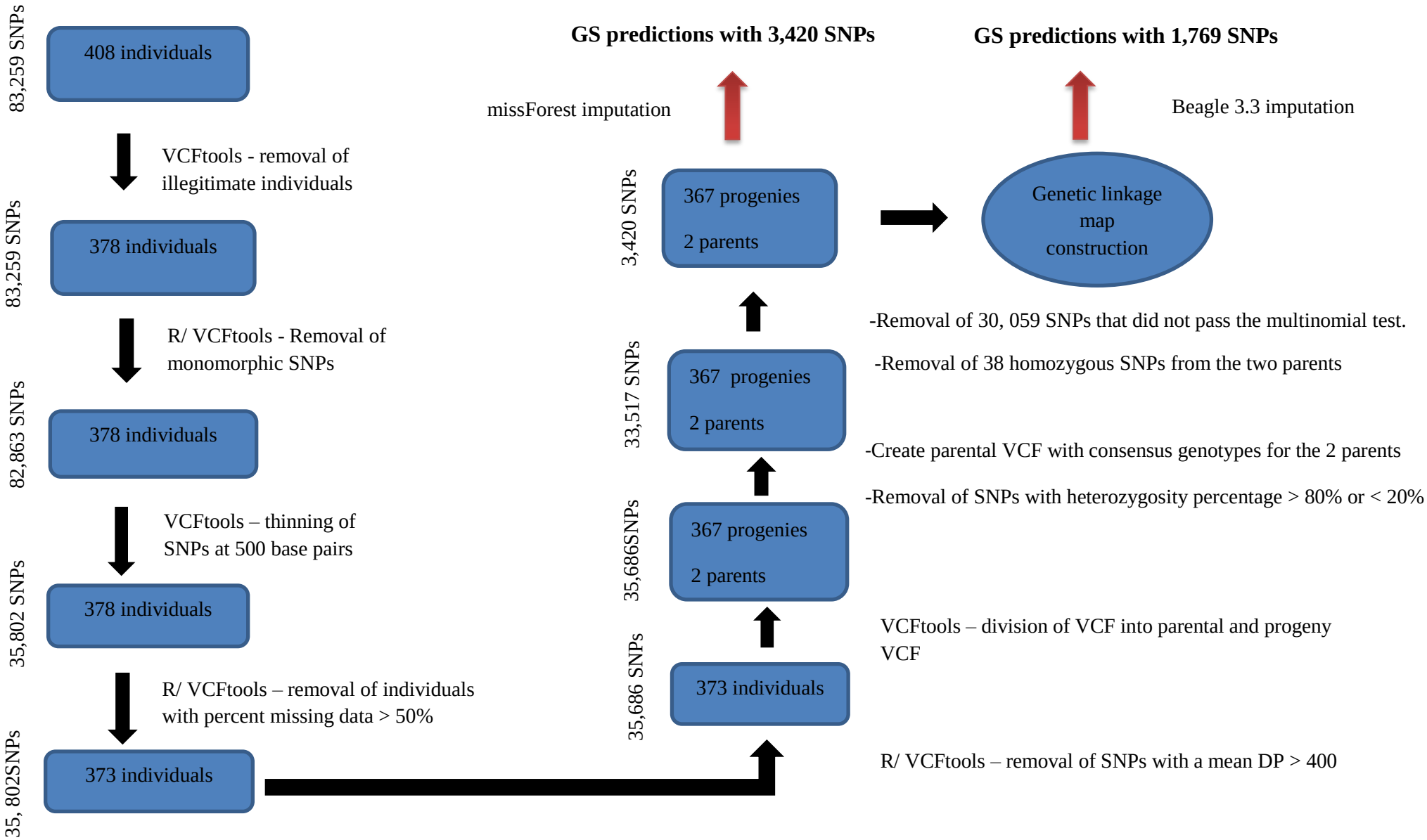


Figure 3. 2: **Pipeline to produce VCF files for genomic predictions**

The pipeline that was developed to check and correct the quality of SNPs and to produce the final VCF file with 3,420 SNPs for use in genomic predictions is shown in Figure 3.2 in the previous page.

3.4 Construction of genetic linkage map

In order to make imputation with Beagle 3.3, a genetic map is required. Prior to linkage map construction, the R software was used to convert the final marker dataset from the VCF file format (0/1, 1/1, 0/0) to the cross-pollination (CP) population type format (*hk*, *kk*, *hh*, *nn*, *ll*, *np*, *lm*) which is compatible with the JoinMap software (Stam, 1993). SNP markers were grouped into two categories based on their segregation patterns. The first group consisted of markers that segregated in a 1:1 ratio and these were the test-cross markers in which one parent was heterozygous whilst the other parent was homozygous (<*nn*×*np*> or <*lm*×*ll*>). The second category of markers segregated in a 1:2:1 ratio, and these were the inter-cross markers in which both parents were heterozygous (<*hk*×*hk*>). To identify linkage groups according to those already in previous studies such as Lespinasse *et al*. (2000) and Pootakham *et al*. (2015), SSR markers were included in the genetic map. Linkage analysis was done using JoinMap 5.0 using parameters set for cross-pollinated (CP) population types. Assignment to linkage groups was done based on the logarithm of the odds (LOD) threshold value of 6.0. The research used linkages with a recombination (REC) rate of < 0.4, a map LOD value of 0.05, and a goodness-of-fit jump threshold set at five for inclusion into the linkage map and for calculating the linear order of markers within a linkage group. The Kosambi mapping function (Kosambi, 1944) was used to estimate map distance and to convert the recombination fractions between markers to map distances in centiMorgans.

3.5 Comparing the performance of marker imputation methods

The Beagle 3.3 (Browning and Browning, 2007) was the map-dependent imputation method used, whilst Random Forest regression algorithm (Stekhoven and Bühlmann, 2012), represented the map independent method. For the two algorithms, imputation was performed on the $p \times n$ matrix of p individuals and n SNP markers whose data points, presented in (0, 1, 2, NA) format, represented the three possible genotypes and the missing value (NA), respectively. Beagle was selected amongst the map dependent imputation methods since it is widely used for GS imputation and because it offers considerable flexibility to tune the imputation algorithm to specific needs depending on the genetic structure of data sets. This is also the imputation software used for the rubber GS study of Cros *et al*. (2019). The Random Forest algorithm was chosen for this research because it came top amongst five other map-independent algorithms in a recent study with GBS data (Rutkoski *et al*., 2013).

3.5.1 The Beagle algorithm

Because a proper physical map for the *Hevea* clones with 18 linkage groups is not available, because of lack of a good reference genome, Beagle 5.1 (the latest version of Beagle) could not be used to impute missing markers since it requires physical distances between SNPs. Beagle 3.3, which is the last version of Beagle able to make imputation from genetic distances without requiring physical distances, was therefore used to perform marker imputation. Imputation was performed on the set of SNPs that were mapped with JoinMap, using the following parameters: *nsamples*=20 and *niterations*=25.

3.5.2 The Random Forest algorithm

For the random forest imputation (RFI), missing SNP markers were estimated using the random forest regression (Breiman, 2001), using all the 3,420 available SNP marker data. The RFI was

implemented in R using the package "missForest" (Stekhoven and Bühlmann, 2012) and the function *missForest*. To decide on the number of iterations and the number of regression trees to grow, a data matrix was created from marker data and random non-missing datapoints (10% of the total) were set as missing using the *ProdNA* function in "missForest" and three combinations (10 iterations and 100 regression trees, 15 iterations and 150 regression tress, 15 iterations and 300 regression trees) of number of iterations and decision trees were used as parameter sizes to impute the data with 10% missing information.

After the imputation, the out-of-bag error (OOB) was computed and the combination of 15 iterations and 300 regression trees was chosen to impute the real data, as it gave the lowest OOB error. To perform the random forest imputation, markers were first converted to a format compatible with RFI as follows: 0/0 as 0, 1/1 as 1, 0/1 or 1/0 as 2 and "."as NA. The RFI procedure used to impute missing SNP datapoints is implemented in missForest as follows:

1. For the marker matrix *M*, SNP markers were first sorted in ascending order (from lowest to highest percent missing) and missing values were then imputed using mean imputation (MNI).
2. At each marker *j* containing missing values, non-missing values were used to grow 300 random forest regression trees ($\boldsymbol{\Theta}1,\ldots,300$). Each of the 300 RF regression trees was grown using a bootstrap sample of individuals *Y*, and a random sample of $\sqrt{n-1}$ marker predictors were used, where *n*-1 is the total number of markers excluding marker *j*. Each of the RF trees ($\boldsymbol{\Theta}$) contains terminal node values and instructions for recursive partitioning of observations into the terminal nodes. These instructions include split variables at each node, and the value of the split variable that is used for partitioning.
3. The missing values at each marker *j* were imputed as shown on equation 3 below:

$$\hat{Y} = \frac{1}{300}\sum_{1}^{300} h(x, \Theta)$$

where x is an input vector.

4. Marker *j* was then updated in the marker matrix *M* using the $\hat{Y}$ values as the estimate of missing values.
5. The steps 2 to 4 were repeated for each marker until all the markers were imputed.
6. Using the imputed matrix, steps 2 to 5 were repeated until convergence occurred or for a maximum of 15 iterations. The convergence was declared as soon as the ΔN went up for the first time as shown on equation 4 below:

$$\Delta N = \frac{\sum_{j \epsilon n}(M_1 - M_0)^2}{\sum_{j \epsilon n}(M_1)^2}$$

Where M_1 is the newly imputed marker matrix and M_0 is the previously imputed marker matrix. When the convergence criteria were met, the research used M_0 as the final estimate of *M*.

Two imputed datasets of SNPs from Beagle and RFI were used to perform genomic predictions, and the accuracy of the two imputation methods was measured by comparing the prediction accuracies obtained with the two imputed SNP datasets.

3.6 Genomic predictions

To perform across-site genomic predictions using rrBLUP, the two marker data files from Beagle and RFI were split into two SNP matrices, the first one contained the 179 HR46 clones and the second SNP matrix contained the 125 Sapest13 clones. To proceed to GS predictions the format of markers was converted to the {-1, 0, 1} format which is compatible with rrBLUP in which 1 is homozygous for allele one, 0 is heterozygous and -1 is homozygous for allele two.

A purely additive GS model and the random regression best linear unbiased prediction (rr-BLUP) method (Meuwissen *et al*., 2001) were used to estimate the genomic estimated genetic values (GEGVs) of the full-sib rubber clones. This was chosen as Cros *et al*. (2019) did not find any difference in accuracy when using this approach and other standard prediction approaches (including approaches modelling non additive effects). Marker effects were estimated using the following linear mixed model:

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

Where *y* is the vector of adjusted phenotypic values, *m* is a vector of the random marker effects, β represent the vector of fixed effects (the mean phenotype), *Z* is the incidence matrix for the vector of random marker effects (*m*) , i.e. the matrix of genotypes, *X* is the incidence matrix for the vector of fixed effects (β), and ε is the vector of residual effects.

The structure of the means and variances for the rr-BLUP model are as follows: $m \sim \mathrm{N}(0,\ \mathrm{G})$, $E(y) = X\beta,\ \varepsilon \sim \mathrm{N}(0,\ \mathrm{R} = I\sigma_\varepsilon^2),\ Var\ (y) = V = ZGZ' + R, \sigma_m^2 =\ (\sigma_a^2/\eta), G =\ I\sigma_m^2$

Where *n* is the total number of marker loci, σ_m^2 is the variance which is common to each marker effect, and σ_e^2 is the residual error variance (Resende *et al*., 2012b). The rr-BLUP mixed model for prediction of *m* is equivalent to:

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + I\frac{\sigma_\varepsilon^2}{(\sigma_a^2/\eta)} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \widehat{m} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix}$$

where η relates to the number of SNP markers used and σ_a^2 and σ_ε^2 refer to the total genetic variance of the trait and the residual variance, respectively. An estimate of the effect of each SNP marker was obtained by solving the mixed model equations presented above. The predicted

genomic estimated genetic value (GEGV) of the individual rubber clone j was given by $\hat{g}_j = \sum Z_{ij}\, \hat{m}_i$. The Z matrix was constructed from the number of alleles that were observed in each marker (-1, 0 or 1).

Residuals are assumed to be normally distributed with zero mean. The genetic variation at each locus, which is the amount of genetic variation explained by each SNP is given by σ_a^2/η, where η relates to the number of SNP markers used, and is given by the equation: $\eta = 2\sum_i^n p_i\,(1 - p_i)$, where p_i represents the allele frequency of one of the alleles of loci i. SNP marker effects *m* and residuals errors *e* are assumed to be independent. The ridge parameter $\lambda = \frac{\sigma e^2}{\sigma m^2}$ was used to control the shrinkage of marker effects.

The variances were calculated by restricted maximum likelihood (REML). The ridge-regression BLUP analysis was performed using the '*mixed.solve*' function in the R software package rrBLUP (Endelman, 2011).

3.7 Across site genomic predictions

Analyses were performed for predictions between sites, leading to two different validation approaches (HR46 towards Sapest13 and Sapest13 towards HR46). In the first scenario, 179 individual clones from HR46 were used to train the GS model, and the 123 clones in Sapest13 were used as the validation population. In the second scenario, the 125 individual clones in Sapest13 were used to train the GS model, to predict the GEGVs of the 177 clones in HR46, that were used as the validation population. The GS predictive ability was obtained for each set as the Pearson correlation between the GEGV ($\hat{g}$) and the phenotype (y) of clones composing the set. The GS accuracy was obtained by dividing the predictive ability by the square root of broad sense heritability (H^2).

3.9 Effect of marker density on GS accuracy

To study the effect of marker density on GS accuracy, the SNP matrices imputed with RFI and Beagle were imported into R. A loop was created in R to select random SNP samples in which each sample size had 30 replicates. The 3,420 markers from Random Forest imputation and the 1,769 markers from Beagle imputation were selected as the first marker subsets, from which SNP markers for lower densities were randomly selected. The different SNP sample sizes for the SNP matrix from RFI were as follows 25 SNPs, 50 SNPs, 100 SNPs, 250 SNPs, 500 SNPs, 1000 SNPs, 2000 SNPs and 3420 SNPs. For Beagle imputation, the different SNP sample sizes were as follows: 25 SNPs, 50 SNPs, 100 SNPs, 250 SNPs, 500 SNPs, 1000 SNPs,1500 SNPs and 1769 SNPs (corresponding to the number of SNPs that could be mapped with JoinMap 5.0). Random sampling of SNPs was done in R using the function *sample*. Genomic predictions were performed at each marker density and for all the 30 replicates, and the GS accuracies of the 30 replicates were computed. The average GS accuracy at each SNP sample size was computed and plots were made to show the effect of marker density on GS accuracy in the two sites and with the two imputation methods.

CHAPTER FOUR

4.0 RESULTS

4.1 To construct a high-density genetic linkage map of rubber clones of a single family.

Of the 3,420 SNP markers in the final VCF file for genomic predictions, 1,769 non-redundant markers could be located on a genetic linkage map that were spread over 18 linkage groups (LG) (Figure 4.1), which almost correspond to the haploid chromosome number of the *Hevea* tree (Lespinasse *et al*., 2000), except for the presence of one extra linkage group (LG6b), indicating that the software split chromosome number 6 into two linkage groups (LG6a and LG6b). Of the 1,769 markers, 1,339 SNP markers were of the segregation type <hk×hk>, 269 SNPs were of the segregation type <lm×ll> and the remaining 161 SNPs were of the segregation type <nn×np>.

The linkage map encompassed 2600.9 cM, with linkage groups ranging from 42.1 cM (LG6b) to 181.8 cM (LG10). The number of unique SNP markers mapped to each linkage group ranged from 20 SNPs in LG6b to 143 SNPs in LG 5, with an average of 98 SNPs per linkage group (Table 4.1). The number of SNPs in each linkage group is also shown on the genetic map below each linkage group. The average inter-marker distance for the linkage map was 1.47 cM, with 61% of the SNP marker intervals less than 1.23 cM. Large gaps of more than 10 cM were observed in LG5, LG6b, LG7, LG13, LG14, LG15, LG16, LG17, and LG18. The distance between adjacent markers was relatively uniformly distributed in LG10, LG11 and LG12 as shown by the maximum marker intervals of 5.26 cM, 5.59 cM and 5.3 cM, respectively.

The linkage map also encompassed 308 SSR markers spread across 18 linkage groups. Of the 308 SSR markers on the genetic linkage map, 42 markers segregated in the <ab×cd> fashion, 88 markers were of the segregation type <ef×eg>, 39 markers in the <hk×hk> pattern, 89 markers in

Table 4. 1: Distribution of SNP markers on the Hevea genetic linkage map

Linkage group	Number of Markers	Length in cM	Average marker interval (cM)	Maximum interval (cM)
LG1	95	130.5	1.37	9.91
LG2	105	131.7	1.25	6.72
LG3	99	152.7	1.54	8.65
LG4	59	137.7	2.3	7.92
LG5	143	147.9	1.03	18.16
LG6a	51	114.1	2.24	9.14
LG6b	20	42.1	2.11	16.65
LG7	78	165.9	2.13	17.29
LG8	84	148.1	1.76	7.87
LG9	112	155.4	1.39	7.17
LG10	139	181.8	1.31	5.26
LG11	120	156.5	1.3	5.59
LG12	86	111.0	1.29	5.30
LG13	108	150.3	1.39	10.04
LG14	124	134.3	1.08	11.84
LG15	112	171.9	1.53	10.56
LG16	70	135.7	1.94	13.56
LG17	68	101.4	1.49	10.23
LG18	96	132.1	1.38	14.55
Total	1769	2600.9		

the <lm×ll> segregation, and 50 markers were of the segregation type <np×np>. The number of unique SSR markers mapped to each of the 18 linkage groups ranged from 02 markers in LG6b to 32 markers in LG10 (Appendix 2). The average SSR inter-marker distance was 8.4 cM.

On the genetic linkage, LG6 represent LG6a, LG7 is LG6b, LG8 is LG7, LG9 is LG8, LG10 is LG9, LG11 is LG10, LG12 is LG11, LG13 is LG12, LG14 is LG13, LG15 is LG14, LG16 is LG15, LG17 is LG16, LG18 is LG17, and LG19 is LG18 according to the identification of linkage groups by Lespinasse *et al*. (2000a)

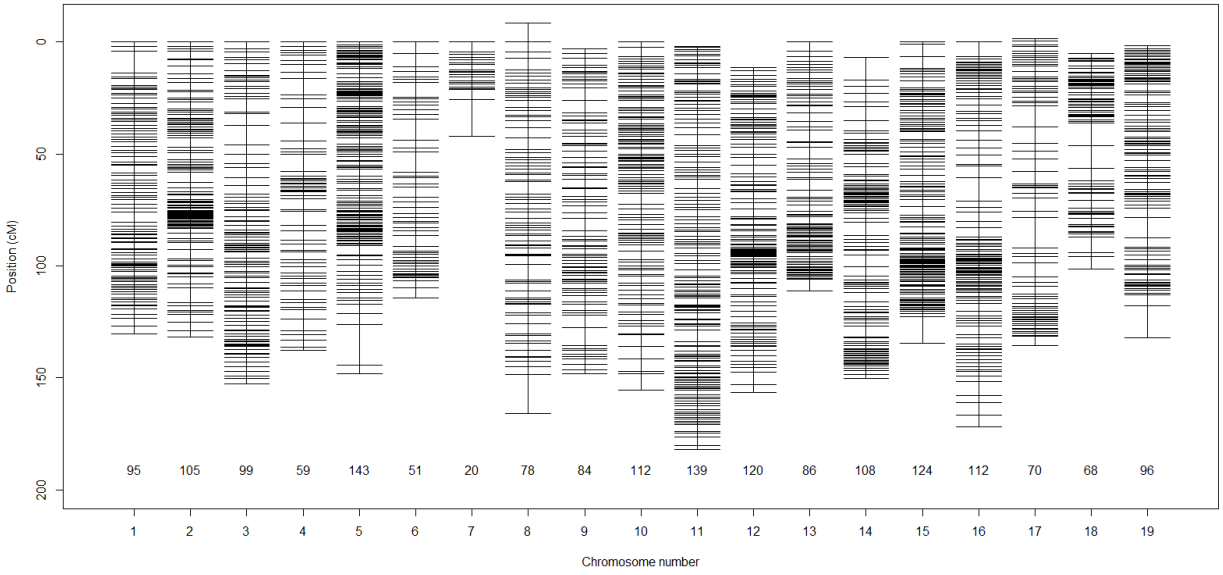


Figure 4. 1: Genetic linkage map of the *Hevea* based on the progeny of the cross PB260 × RRIM600

4.2 To compare the performance of marker imputation methods on GS accuracy

Results from comparing the two imputation methods showed that Beagle 3.3 performed better than random forest imputation. With all markers, and despite the fact that Beagle could only be used on the subset of mapped SNPs, the average GS prediction accuracy over the two sites

obtained with Beagle was 0.52, whilst with random forest imputed marker data it was only 0.48. In addition, Beagle performed better than random forest imputation in both HR46 and Sapest13. When the performance of HR46 clones was predicted using Sapest13 clones as the training population, molecular marker data from Beagle gave an accuracy of 0.54 whilst marker data from random forest imputation gave an accuracy of 0.5. Similarly, when the yield performance of rubber clones in Sapest13 was predicted using HR46 clones as the training population, Beagle imputed marker data gave a GS accuracy of 0.5 whilst marker data from random forest imputation gave a GS accuracy of 0.45.

Across-site GS predictions using data from the two imputation methods showed that GS accuracy was higher when the 125 rubber clones in Sapest13 were used as training population to predict GEGVs of the 177 clones in HR46 than when the 179 rubber clones in HR46 were used to train the GS model to predict the GEGVs of the 123 clones in Sapest13, despite the larger training population in HR46 as compared to Sapest13. The between site difference in GS accuracy when performing Sapest13 towards HR46 predictions and HR46 towards Sapest13 predictions with marker data from the two imputation methods was almost the same, 0.05 for Beagle and 0.04 for Random Forest imputation.

In addition, the saturation point at which increasing marker density resulted in a very small increase in GS accuracy was different for the two imputation methods. Despite fewer number of markers (1,769) of Beagle imputation as compared to random Forest imputation, a plateau was reached with fewer markers (1,000) with Beagle imputation data whilst with marker data from random forest imputation, a plateau was reached at 2,000 markers in both directions of prediction (Figure 4.2).

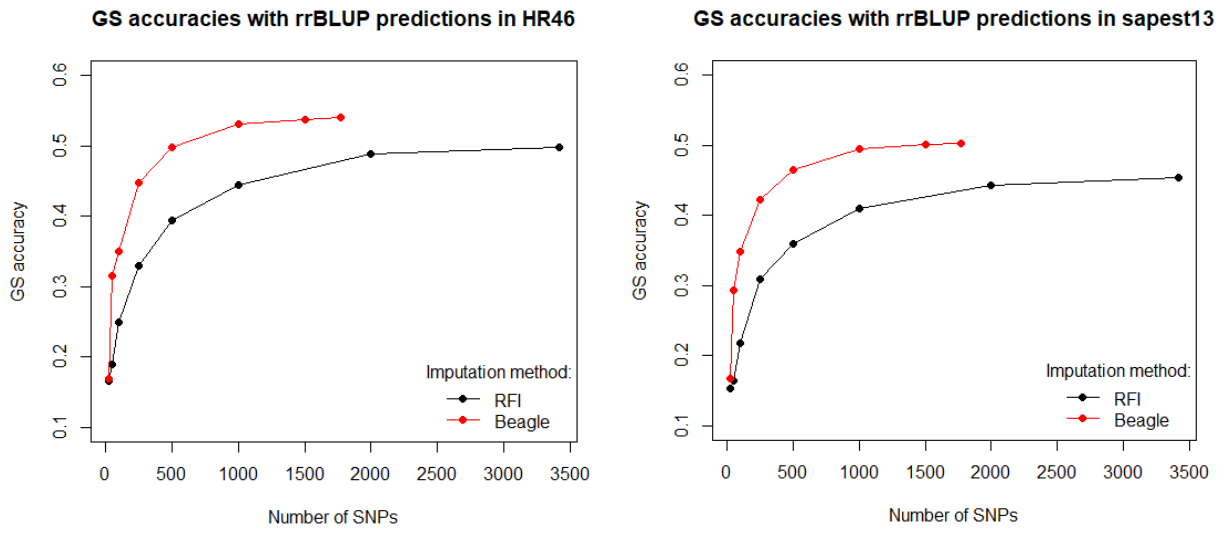


Figure 4. 2: Effect of imputation approach on genomic selection accuracy in HR46 (left) and Sapest13 (right). When not all markers were used, values are means over 30 replicates

4.3 Effect of Marker density on GS accuracy

Results from across-site GS predictions in the two directions, that is, from Sapest13 towards HR46 and HR46 towards Sapest13, showed that GS accuracy increased with increasing marker density. By performing GS predictions from Sapest13 towards HR46, the increase in marker density from 25 SNPs to the total number of SNPs in respective imputation methods resulted in an increase in GS accuracy. With a marker density of 25, 50, 100, 250, 500, 1000, 1500, and 1769 with Beagle imputation marker data, the GS accuracy increased from 0.17, 0.32, 0.35, 0.45, 0.5, 0.53, 0.54, and 0.54, respectively (Table 4.3). Performing GS predictions from Sapest13 towards HR46 using marker data from Random Forest imputation, and with marker densities of 25, 50, 100, 250, 500, 1000, 2000, and 3420 SNP markers resulted in an increase in GS accuracy of 0.17, 0.19, 0.25, 0.33, 0.39, 0.44, 0.49, and 0.5, respectively (Table 4.2). On the other hand,

performing GS predictions from HR46 towards Sapest13 with Beagle imputation marker data, marker densities of 25, 50, 100, 250, 500, 1000, 1500, and 1769 SNPs resulted in GS accuracies of 0.17, 0.19, 0.25, 0.33, 0.39, 0.44, 0.48, and 0.5, respectively.

Table 4. 2: Genomic selection accuracy according to marker density using random forest imputed marker data. When not all markers were used, values are means over 30 replicates

Number of SNPs	Sapest13 GS accuracies	HR46 GS accuracies
25	0.15	0.17
50	0.16	0.19
100	0.22	0.25
250	0.31	0.33
500	0.36	0.39
1000	0.41	0.44
2000	0.44	0.49
3420	0.45	0.5

Performing HR46 towards Sapest13 GS predictions with marker data from Random Forest imputation using SNP marker densities of 25, 50, 100, 250, 500, 1000, 2000, and 3420 resulted in an increase in GS accuracy of 0.15, 0.16, 0.22, 0.31, 0.36, 0.41, 0.44, and 0.45, respectively. In addition, by performing Sapest13 towards HR46 GS predictions using marker data from Beagle imputation, increasing marker density from 25 SNPs to 500 SNPs resulted in a sharp increase in GS accuracy from 0.17 with 25 markers to 0.5 with 500 markers. The increase in GS accuracy with increasing marker density started to decline when markers were increased from

500 SNPs to 1000 SNPs as shown in Figure 4.2. Beyond 1000 markers up to the 1769 markers of Beagle imputation, the increase in GS accuracy in response to increasing marker density started to plateau (0.53 to 0.54), indicating that a point will be reached at which increasing marker density will result in no significant increase in GS accuracy.

Table 4. 3: Genomic selection accuracy according to marker density using Beagle 3.3 imputed marker data. When not all markers were used, values are means over 30 replicates.

Number of SNPs	Sapest13 GS accuracies	HR46 GS accuracies
25	0.17	0.17
50	0.29	0.32
100	0.35	0.35
250	0.42	0.45
500	0.46	0.5
1000	0.49	0.53
1500	0.5	0.54
1769	0.5	0.54

The same trend was seen when HR46 towards Sapest13 GS predictions were performed, except for the differences in accuracy between the two prediction directions in which an increase in marker density from 25 to 500 SNPs also led to a sharp increase in GS accuracy from 0.17 to 0.46. From 1000 markers up to the total number of markers (1769), the GS accuracy only increased from 0.49 to 0.5.The trend in which a marker saturation point is reached was also observed when molecular marker data from random forest imputation was used to perform GS

predictions across-sites (Sapest13 towards HR46 and HR46 towards Sapest13). By performing Sapest13 towards HR46 predictions, increasing marker density from 25 to 2000 resulted in an increase in GS accuracy from 0.17 to 0.49. However, beyond 2000 markers, a plateau was reached in which increasing marker density from 2000 to 3420 markers only resulted in a GS accuracy increase from 0.49 to 0.5. When GS predictions were performed from HR46 towards Sapest13 using random forest imputed data, there was a sharp increase in GS accuracy from 0.15 to 0.44 when marker density was increased from 25 markers to 2000 markers. As in Sapest13 towards HR46 predictions, beyond 2000 markers the increase in GS accuracy in response to increasing marker density started to stagnate as shown by a small increase in GS accuracy from 0.44 to 0.45 (Table 4.2 and Table 4.3).

CHAPTER FIVE

5.0 DISCUSSION

For plant breeders to effectively perform genomics assisted breeding, there is need for a thorough understanding of the underlying factors affecting the accuracy of genomic selection. The present study utilized a panel of 304 rubber clones that were genotyped with 3420 SNP markers to study the effect of marker imputation approach and marker density on genomic selection accuracy, among other objectives. The findings of this research will assist plant breeders to optimize their breeding programs, thus allowing them to make the most effective and efficient use of resources when implementing genomics-assisted breeding.

5.1 Constructing a high-density genetic linkage map of rubber clones of a single family

The use of falsely discovered markers in genetic linkage map construction could result in a linkage map of compromised quality. Thus, to obtain a high quality *Hevea* linkage map, stringent criteria were applied to call and filter SNP markers. To obtain high quality markers for linkage map construction, the pattern of allelic segregation for chi-square goodness-of-fit to the expected Mendelian segregation ratios was tested for each locus, and all markers with significant segregation distortion were discarded from further analysis in the study. Although the exclusion of SNPs with significant segregation distortion from linkage map construction results in low marker coverage of the genome, these markers were not included in linkage map construction to ensure production of a high-quality map. Markers which showed a deviation from the expected Mendelian segregation ratios, that is, with significant segregation distortion, could have led to the underestimation or overestimation of recombinant fractions, which ultimately affects both the calculation of genetic distances between markers and the order of markers (Zhao *et al*., 2018).

The high proportion of SNP markers that could not be mapped (1,651 out of 3,420) reveals the high proportion of genotyping errors on these markers, which is the case especially with SNPs obtained by GBS.

A genetic linkage map with 1,769 non-redundant SNP markers spread over 18 linkage groups in which LG6 was split into LG6a and LG6b was obtained. The length of the linkage map (2600.9 cM) is comparable to previously published linkage maps in rubber and other crops (2,144 cM in Lespinasse *et al*. (2000b), 2,041 cM in Pootakham *et al*. (2015) and 2,441 cM in Le Guen *et al*. (2011)). The close similarity between the linkage map obtained in this research and the one obtained by Le Guen *et al*. (2011) could be because of a common parent (PB260) between the two populations used. The map had an extra linkage group (LG6b) which resulted in a linkage map with one LG in excess of the haploid chromosome number of *Hevea* (*n* = 18) (Leitch *et al*., 1998). A *Hevea* genetic map with 19 linkage groups has been reported by Pootakham *et al*. (2015). The reason for an extra linkage group could be due to a large interval without a marker, making the SNP alleles on the two chromosome segments appear as statistically unrelated. This could also be because of a recombination hotspot for which it is difficult to prove the genetic proximity between physically close markers.

The GBS-based linkage map showed significantly higher marker density as compared to previously published microsatellite-based linkage maps obtained using populations of similar or smaller size (one marker in every 8 cM in Le Guen *et al*. (2011), and one marker in every 10 cM in Souza *et al*. (2013)). The marker density of the obtained GBS-based linkage map is comparable to the map obtained by (Pootakham et al., 2015). The higher marker densities in GBS-based linkage maps is primarily due to the genotyping technique, which was defined with

the goal of generating large amounts of markers, as today high marker densities are required for many genomic applications, and in particular genomic selection (Pootakham *et al*., 2015).

Although SNP markers were relatively uniformly distributed on the linkage map, nine gaps larger than 10 cM were observed. The largest gap was 18.16 cM, and was located at the terminal portion of LG 5. The number of large gaps greater than 10 cM and the largest inter-marker distance in the linkage map is comparable to the linkage map of Pootakham *et al*. (2015) which had seven gaps greater than 10 cM and a maximum inter-marker distance of 21 cM. The presence of these gaps could be as a result of the limitation of GBS-SNP markers in detecting polymorphisms in certain regions of the genome. Apart from the above, these large gaps could represent recombination hotspots or genome sections that were identical-by-descent among the two parents and thus lack polymorphisms. The large gaps exhibiting a low degree of polymorphisms have also been observed in genetic linkage maps of rubber and other crop species, such as rye and common bean (Galeceae *et al*., 2011).

The genetic map also encompassed 308 SSR markers spread between 18 linkage groups. The SSR inter-marker distance (one marker in every 8.4 cM) in the constructed genetic linkage map is almost similar to the marker density obtained in the microsatellite-based linkage map of Le Guen *et al*. (2011) (one marker in every 8 cM). The close similarity in microsatellite marker densities between the two maps could be because of the common parent (PB260) shared between the clones used in the two studies. It is thanks to these SSR markers that it was possible to identify the linkage groups according to those already defined in previous studies. It is also thanks to these 308 SSR markers that we were able to know that it is linkage group 6 (LG6) which was split into two fragments.

The linkage map will serve as an important tool to rubber breeders for genomics-assisted breeding, QTL analyses and also to better assemble the *Hevea* whole-genome sequence.

5.2 To compare the performance of Beagle 3.3 and random forest imputation

By comparing the effect of marker imputation approaches on GS accuracy, a relatively higher GS accuracy was obtained with Beagle 3.3 imputed marker data as compared to random forest imputation. Beagle imputed marker data gave an average GS accuracy of 0.52 using 1,769 markers whilst random forest imputation gave an average GS accuracy of 0.48 using 3,420 markers, which is almost double the number of markers in Beagle. The higher accuracy in Beagle imputation as compared to random forest imputation could be attributed to the use of linkage or haplotype information in Beagle since it is a map-dependent imputation approach (He *et al*., 2015). Random forest imputation is a map independent imputation, and hence it does not use haplotype information to impute missing markers.

The results obtained with Beagle imputation are useful especially for orphan species that do not have a proper reference genome. With orphan species, making a genetic linkage map with SNP markers is a good option for performing map-based imputation approaches, because Beagle imputation performed better than random forest imputation. These findings clearly demonstrate the importance of the availability of a high-quality genetic map for orphan species without a proper reference genome in order to perform Beagle imputation.

However, since the difference in GS accuracy between random Forest imputation and Beagle imputation is not that big, Random Forest imputation is still a satisfactory option so scientists can make genomic predictions with markers imputed with Random Forest imputation in case a genetic map cannot be constructed because of the population which is not suitable for that.

5.3 Effect of Marker density on genomic selection accuracy

The effect of marker density on GS prediction accuracy was assessed by making random SNP samples of varying sizes in which SNP data sets from Beagle imputation (1769 SNPs) and Random Forest imputation (3420 SNPs) were used as the source SNP data sets from which random SNP sampling was made. There was a strong response in GS accuracy to increasing marker density with marker data from both Beagle and RFI. In Beagle imputation, there was a sharp increase in GS accuracy as the marker density was increased from 25 to 1000 markers, as shown by an increase in GS accuracy from 0.17 to 0.53 when prediction was done using Sapest13 clones as the training population and from 0.17 to 0.44 when prediction was done using Sapest13 clones as the selection population. The sharp increase in GS accuracy was also observed with marker data from RFI in which there was a sharp increase in GS from 0.17 to 0.49 as the marker density was increased from 25 SNPs to 2000 SNPs when Sapest13 clones were used as the training population and from 0.15 to 0.44 when Sapest13 clones were used to train the GS model. The observed sharp increase in GS accuracy as a result of an increase in marker density is because at a very low marker density (25 SNPs), the probability of linkage disequilibrium between the QTLs and markers is very low and hence only a smaller proportion of genetic variation is explained. However, as the number of SNPs is increased to 1000 markers, most QTLs will be in LD with some genetic markers and estimates of marker effects will lead to accurate predictions of clone genetic values (Ala Noshahr *et al*., 2018). For example, using the linkage map developed in this research with 2600.9 cM and with a marker density of 25 SNPs, on average there is one marker for every 104 cM, but when all the 1769 markers of Beagle imputation are used there is one marker in every 1.47 cM. Results of this study are similar to

those of Norman *et al*. (2018) and Liu *et al*. (2015) who also reported an increase in GS accuracy in response to an increase in marker density.

Results of this study also showed that a plateau was reached at which increasing marker density resulted in a very small or no increase in GS accuracy. For Beagle and Random Forest imputation, a very small increase in GS accuracy was observed beyond 1000 and 2000 markers, respectively, indicating that there is no reason for going beyond this plateau as that would increase genotyping costs without any benefit to GS accuracy. Because the plateau was reached at a much lower marker density in Beagle as compared to RFI (half the marker density of RFI), and with a higher GS accuracy, it shows that Beagle imputation would be the best option for plant breeders. Results obtained with marker data from Beagle imputation are similar to those of Nielsen *et al*. (2016) who reported that a minimum marker set of 1,000 is required to avoid a decline in prediction accuracy. These SNP numbers are significantly lower than the plateau point of previous studies (Norman et al., 2018). The lower number of SNPs required to reach the plateau could be attributed to the high relatedness between the training and selection population since the study used within family rubber clones. This is because, the more related the training and selection population is, the lower the effective number of independent loci (*Me*), and the higher the accuracy of genomic selection (Daetwyler *et al*., 2013). The lower number of markers needed to reach a plateau was due to a high LD which was brought about by a low *Me* (Meuwissen *et al*., 2001). This explanation is supported by Heff et al. (2011) who reported a significant reduction in genomic selection model accuracy after using information from half-sibs instead of full-sibs, indicating the importance of relatedness between the training and validation population. In addition, the lower marker density required to reach a plateau in this research as compared to the research of Norman et al. (2018b) could also be attributed to the low effective

population size (*Ne*) in within family rubber clones as compared to the wheat lines studied by the previous author. The *Ne* determines the threshold at which a plateau is reached through its effect on LD. The LD between QTLs and markers in the genome determines the amount of markers needed to reach a plateau (Wang et al., 2017). Generally, at a low *Ne*, as in within family individuals, the number of independent segments in the genome is expected to be small, and fewer independent segments means that fewer markers are needed to mark all segments. Here, the research used within family rubber clones, which means they have a low *Ne*, and hence a higher GS accuracy at a lower marker density.

The reason for reaching a plateau in which increasing marker density resulted in a very small or no increase in GS accuracy could be because almost all QTLs were in LD with some genetic markers. Thus, at the ceiling point, the most important markers were in LD with the QTLs for rubber yield and therefore increasing marker density beyond the ceiling point only added markers that have very small or no effect to the trait. In addition this could be because as the number of SNPs increases, more SNPs will start to support the same QTL (Habier *et al*., 2011), and hence a plateau is reached.

5.4 Comparison with published results using the same individuals and phenotypic data

The mean between site GS accuracy obtained here with all markers (0.52) is very close to the one of Cros *et al*. (2019), who performed GS in rubber using SSR markers and using the same phenotypic data used in this research: they reported a GS accuracy of 0.53 using a set of 332 SSRs and 330 rubber clones (which is 28 more clones than used in this research) showing that SNPs from GBS, despite the fact that they have higher percentage of missing data and higher error rate than SSRs, can be used for GS predictions with no loss in accuracy. In addition, it opens the way to the practical application of GS. Indeed, Cros *et al*. (2019) showed that, to

increase the genetic gain in the studied rubber cross, GS required to be applied on a large number of selection candidates (>1000), which would not be cost effective with SSRs. However, this becomes feasible with GBS, thanks to its much lower cost per sample.

CHAPTER SIX

6.0 CONCLUSION AND RECOMMENDATIONS

6.1 Conclusion

This study demonstrates that GBS is a rapid, efficient and cost-effective approach for performing genomics-assisted breeding and construction of a high-density genetic linkage map for *Hevea*. Since GBS is a highly reproducible genome sequencing technique, other plant scientists can employ the same protocol to construct linkage maps for *Hevea* using different mapping populations based on the same set of SNPs. Results of the present study demonstrate that genomic selection has huge potential to increase rubber latex yield and to reduce the generation interval in *Hevea* through increasing genetic gain. In addition, Beagle imputation proved to be the marker imputation approach of choice compared to Random Forest imputation when performing genomics-assisted breeding. For species in which it is difficult to find a proper genetic linkage map, Random Forest imputation could still be a good option for marker imputation. Finally, results of this study showed that the minimum marker density required to achieve a good accuracy is determined by the relatedness between the training and selection population and also the marker imputation approach.

6.2 Recommendations

- Further research should consider other traits of interest for rubber, like vegetative growth, resistance to diseases, etc.
- More research comparing genomic selection prediction mathematical models using large training and selection populations is needed in order to have a thorough understanding of factors that affect the accuracy of genomic selection.

- Further research should be done to model genotype by environment interactions in genomic selection.

REFERENCES

Akdemir, D., Isidro-Sánchez, J., 2019. Design of training populations for selective phenotyping in genomic prediction. Sci. Rep. 9, 1446. https://doi.org/10.1038/s41598-018-38081-6

Ala Noshahr, F., Rafat, S., Imany Nabiyyi, R., Alijani, S., Robert-Granié, C., 2018. The Impact of Different Genetic Architectures on Accuracy of Genomic Selection Using Three Bayesian Methods. Iran. J. Appl. Anim. Sci. 8, 53–59.

Alipour, H., Bai, G., Zhang, G., Bihamta, M.R., Mohammadi, V., Peyghambari, S.A., 2019. Imputation accuracy of wheat genotyping-by-sequencing (GBS) data using barley and wheat genome references. PLoS ONE 14. https://doi.org/10.1371/journal.pone.0208614

Annicchiarico, P., Nazzicari, N., Li, X., Wei, Y., Pecetti, L., Brummer, E.C., 2015. Accuracy of genomic selection for alfalfa biomass yield in different reference populations. BMC Genomics 16, 1020. https://doi.org/10.1186/s12864-015-2212-y

Atefi, A., Shadparvar, A.A., Hossein-Zadeh, N.G., Atefi, A., Shadparvar, A.A., Hossein-Zadeh, N.G., 2016. Comparison of whole genome prediction accuracy across generations using parametric and semi parametric methods. Acta Sci. Anim. Sci. 38, 447–453. https://doi.org/10.4025/actascianimsci.v38i4.32023

Aurélien, M., Monteuuis, O., 2017. Rubber tree clonal plantations: Grafted vs self-rooted plant material. Bois Forets Trop. 57–68.

Baird, N.A., Etter, P.D., Atwood, T.S., Currey, M.C., Shiver, A.L., Lewis, Z.A., Selker, E.U., Cresko, W.A., Johnson, E.A., 2008. Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. PLOS ONE 3, e3376. https://doi.org/10.1371/journal.pone.0003376

Bandyopadhyay, S., Agrawal, S.L., Ameta, R., Dasgupta, S., Mukhopadhyay, R., Deuri, A.S., Ameta, S.C., Ameta, Rakshit, 2008. An Overview of Rubber Recycling. Prog. Rubber Plast. Recycl. Technol. 24, 73–112. https://doi.org/10.1177/147776060802400201

Bartholomé, J., Van Heerwaarden, J., Isik, F., Boury, C., Vidal, M., Plomion, C., Bouffier, L., 2016. Performance of genomic prediction within and across generations in maritime pine. BMC Genomics 17, 604. https://doi.org/10.1186/s12864-016-2879-8

Beaulieu, J., Doerksen, T., MacKay, J., Rainville, A., Bousquet, J., 2014a. Genomic selection accuracies within and between environments and small breeding groups in white spruce. BMC Genomics 15, 1048.

Belgiu, M., Drăguţ, L., 2016. Random forest in remote sensing: A review of applications and future directions. ISPRS J. Photogramm. Remote Sens. 114, 24–31. https://doi.org/10.1016/j.isprsjprs.2016.01.011

Bernardo, R., Yu, J., 2007. Prospects for Genomewide Selection for Quantitative Traits in Maize. Crop Sci. 47, 1082–1090. https://doi.org/10.2135/cropsci2006.11.0690

Berthelot, K., Lecomte, S., Estevez, Y., Peruch, F., 2014. Hevea brasiliensis REF (Hev b 1) and SRPP (Hev b 3): An overview on rubber particle proteins. Biochimie 106, 1–9. https://doi.org/10.1016/j.biochi.2014.07.002

Beyene, Y., Semagn, K., Mugo, S., Tarekegne, A., Babu, R., Meisel, B., Sehabiague, P., Makumbi, D., Magorokosho, C., Oikeh, S., Gakunga, J., Vargas, M., Olsen, M., Prasanna, B.M., Banziger, M., Crossa, J., 2015. Genetic Gains in Grain Yield Through Genomic Selection in Eight Bi-parental Maize Populations under Drought Stress. Crop Sci. 55, 154. https://doi.org/10.2135/cropsci2014.07.0460

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., Prabhu, K.V., 2016. Genomic Selection in the Era of Next Generation Sequencing for Complex Traits in Plant Breeding. Front. Genet. 7. https://doi.org/10.3389/fgene.2016.00221

Bhatia, D., Wing, R., Singh, K., 2013. Genotyping by sequencing, its implications and benefits. Crop Improv. 40, 101–111.

Biau, G., Scornet, E., 2016. A random forest guided tour. TEST 25, 197–227. https://doi.org/10.1007/s11749-016-0481-7

Blonk, R.J.W., Komen, H., Kamstra, A., Arendonk, J.A.M. van, 2010. Estimating Breeding Values With Molecular Relatedness and Reconstructed Pedigrees in Natural Mating Populations of Common Sole, Solea Solea. Genetics 184, 213–219. https://doi.org/10.1534/genetics.109.110536

Breiman, L., 2001. Random Forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324

Browning, B.L., Zhou, Y., Browning, S.R., 2018. A One-Penny Imputed Genome from Next-Generation Reference Panels. Am. J. Hum. Genet. 103, 338–348. https://doi.org/10.1016/j.ajhg.2018.07.015

Browning, S.R., Browning, B.L., 2007. Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering. Am. J. Hum. Genet. 81, 1084–1097. https://doi.org/10.1086/521987

Calle, Z., Schlumpberger, B.O., Piedrahita, L., Leftin, A., Hammer, S.A., Tye, A., Borchert, R., 2010. Seasonal variation in daily insolation induces synchronous bud break and flowering in the tropics. Trees 24, 865–877. https://doi.org/10.1007/s00468-010-0456-3

Cao, S., Loladze, A., Yuan, Y., Wu, Y., Zhang, A., Chen, J., Huestis, G.M., Cao, J., Chaikam, V., Olsen, M., Prasanna, B.M., Vicente, F.S., Zhang, X., 2017. Genome-Wide Analysis of Tar Spot Complex Resistance in Maize Using Genotyping-by-Sequencing SNPs and Whole-Genome Prediction. Plant Genome 10. https://doi.org/10.3835/plantgenome2016.10.0099

Carron, M.P., Lardet, L., Leconte, A., Dea, B.G., Keli, J., Granet, F., Julien, J., Teerawatanasuk, K., Montoro, P., 2009. FIELD TRIALS NETWORK EMPHASIZES THE IMPROVEMENT OF GROWTH AND YIELD THROUGH MICROPROPAGATION IN RUBBER TREE (HEVEA BRASILIENSIS, MUËLL.-ARG.). Acta Hortic. 485–492. https://doi.org/10.17660/ActaHortic.2009.812.70

Cericola, F., Lenk, I., Fè, D., Byrne, S., Jensen, C.S., Pedersen, M.G., Asp, T., Jensen, J., Janss, L., 2018. Optimized Use of Low-Depth Genotyping-by-Sequencing for Genomic

Prediction Among Multi-Parental Family Pools and Single Plants in Perennial Ryegrass (Lolium perenne L.). Front. Plant Sci. 9. https://doi.org/10.3389/fpls.2018.00369

Chairungsee, N., Gay, F., Thaler, P., Kasemsap, P., Thanisawanyangkura, S., Chantuma, A., Jourdan, C., 2013. Impact of tapping and soil water status on fine root dynamics in a rubber tree plantation in Thailand. Front. Plant Sci. 4. https://doi.org/10.3389/fpls.2013.00538

Chan, A.W., Hamblin, M.T., Jannink, J.-L., 2016. Evaluating Imputation Algorithms for Low-Depth Genotyping-By-Sequencing (GBS) Data. PLOS ONE 11, e0160733. https://doi.org/10.1371/journal.pone.0160733

Chandrasekhar, T.R., Alice, J., Varghese, Y.A., Saraswathyamma, C.K., Vijayakumar, K.R., 2005. GIRTH GROWTH OF RUBBER (HEVEA BRASILIENSIS) TREES DURING THE IMMATURE PHASE. J. Trop. For. Sci. 17, 399–415.

Clark, S.A., Hickey, J.M., Daetwyler, H.D., van der Werf, J.H., 2012. The importance of information on relatives for the prediction of genomic breeding values and the implications for the makeup of reference data sets in livestock breeding schemes. Genet. Sel. Evol. GSE 44, 4. https://doi.org/10.1186/1297-9686-44-4

Cobb, J.N., Juma, R.U., Biswas, P.S., Arbelaez, J.D., Rutkoski, J., Atlin, G., Hagen, T., Quinn, M., Ng, E.H., 2019. Enhancing the rate of genetic gain in public-sector plant breeding programs: lessons from the breeder's equation. Theor. Appl. Genet. 132, 627–645. https://doi.org/10.1007/s00122-019-03317-0

Costa, R.B. da, Resende, M.D.V. de, Araújo, A.J. de, Gonçalves, P. de S., Silva, M. de A., 2000. Maximization of genetic gain in rubber tree (Hevea) breeding with effective size restriction. Genet. Mol. Biol. 23, 457–462. https://doi.org/10.1590/S1415-47572000000200035

Cros, D., Bocs, S., Riou, V., Ortega-Abboud, E., Tisné, S., Argout, X., Pomiès, V., Nodichao, L., Lubis, Z., Cochard, B., Durand-Gasselin, T., 2017. Genomic preselection with

genotyping-by-sequencing increases performance of commercial oil palm hybrid crosses. BMC Genomics 18, 839. https://doi.org/10.1186/s12864-017-4179-3

Cros, D., Denis, M., Sánchez, L., Cochard, B., Flori, A., Durand-Gasselin, T., Nouy, B., Omoré, A., Pomiès, V., Riou, V., Suryana, E., Bouvet, J.-M., 2015. Genomic selection prediction accuracy in a perennial crop: case study of oil palm (Elaeis guineensis Jacq.). Theor. Appl. Genet. 128, 397–410. https://doi.org/10.1007/s00122-014-2439-z

Cros, D., Mbo-Nkoulou, L., Bell, J.M., Oum, J., Masson, A., Soumahoro, M., Tran, D.M., Achour, Z., Le Guen, V., Clement-Demange, A., 2019. Within-family genomic selection in rubber tree (Hevea brasiliensis) increases genetic gain for rubber production. Ind. Crops Prod. 138, 111464. https://doi.org/10.1016/j.indcrop.2019.111464

Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., de los Campos, G., Burgueño, J., González-Camacho, J.M., Pérez-Elizalde, S., Beyene, Y., Dreisigacker, S., Singh, R., Zhang, X., Gowda, M., Roorkiwal, M., Rutkoski, J., Varshney, R.K., 2017. Genomic Selection in Plant Breeding: Methods, Models, and Perspectives. Trends Plant Sci. 22, 961–975. https://doi.org/10.1016/j.tplants.2017.08.011

Daetwyler, H.D., Calus, M.P.L., Pong-Wong, R., Campos, G. de los, Hickey, J.M., 2013. Genomic Prediction in Animals and Plants: Simulation of Data, Validation, Reporting, and Benchmarking. Genetics 193, 347–365. https://doi.org/10.1534/genetics.112.147983

Daetwyler, H.D., Kemper, K.E., van der Werf, J.H.J., Hayes, B.J., 2012. Components of the accuracy of genomic prediction in a multi-breed sheep population. J. Anim. Sci. 90, 3375–3384. https://doi.org/10.2527/jas.2011-4557

Daetwyler, H.D., Pong-Wong, R., Villanueva, B., Woolliams, J.A., 2010. The Impact of Genetic Architecture on Genome-Wide Evaluation Methods. Genetics 185, 1021–1031. https://doi.org/10.1534/genetics.110.116855

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., McVean, G., Durbin, R., 2011. The variant call

format and VCFtools. Bioinformatics 27, 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

Dassonneville, R., Brøndum, R.F., Druet, T., Fritz, S., Guillaume, F., Guldbrandtsen, B., Lund, M.S., Ducrocq, V., Su, G., 2011. Effect of imputing markers from a low-density chip on the reliability of genomic breeding values in Holstein populations. J. Dairy Sci. 94, 3679–3686. https://doi.org/10.3168/jds.2011-4299

Daud, N.W., Mokhatar, S.J., Ishak, C.F., 2012. Assessment of selected Hevea brasiliensis (RRIM 2000 Series) seeds for rootstocks production. https://doi.org/10.5897/AJAR12.272

De Almeida Filho, J.E., Guimarães, J.F.R., e Silva, F.F., de Resende, M.D.V., Muñoz, P., Kirst, M., Resende Jr, M.F.R., 2016. The contribution of dominance to phenotype prediction in a pine breeding and simulated population. Heredity 117, 33–41. https://doi.org/10.1038/hdy.2016.23

De los Campos, G., Hickey, J.M., Pong-Wong, R., Daetwyler, H.D., Calus, M.P.L., 2013. Whole-Genome Regression and Prediction Methods Applied to Plant and Animal Breeding. Genetics 193, 327–345. https://doi.org/10.1534/genetics.112.143313

De los Campos, G., Naya, H., Gianola, D., Crossa, J., Legarra, A., Manfredi, E., Weigel, K., Cotes, J.M., 2009. Predicting Quantitative Traits With Regression Models for Dense Molecular Markers and Pedigree. Genetics 182, 375–385. https://doi.org/10.1534/genetics.109.101501

De Souza, Livia M., dos Santos, L.H.B., Rosa, J.R.B.F., da Silva, C.C., Mantello, C.C., Conson, A.R.O., Scaloppi, E.J., Fialho, J. de F., de Moraes, M.L.T., Gonçalves, P. de S., Margarido, G.R.A., Garcia, A.A.F., Le Guen, V., de Souza, A.P., 2018. Linkage Disequilibrium and Population Structure in Wild and Cultivated Populations of Rubber Tree (Hevea brasiliensis). Front. Plant Sci. 9. https://doi.org/10.3389/fpls.2018.00815

Dimitrijevic, A., Horn, R., 2018. Sunflower Hybrid Breeding: From Markers to Genomic Selection. Front. Plant Sci. 8, 2238. https://doi.org/10.3389/fpls.2017.02238

Dornelas, M.C., Rodriguez, A.P.M., 2005. The rubber tree (Hevea brasiliensis Muell. Arg.) homologue of the LEAFY/FLORICAULA gene is preferentially expressed in both male and female floral meristems. J. Exp. Bot. 56, 1965–1974. https://doi.org/10.1093/jxb/eri194

Duangjit, J., Causse, M., Sauvage, C., 2016. Efficiency of genomic selection for tomato fruit quality. Mol. Breed. 36, 29. https://doi.org/10.1007/s11032-016-0453-3

Edriss, V., Gao, Y., Zhang, X., Jumbo, M.B., Makumbi, D., Olsen, M.S., Crossa, J., Packard, K.C., Jannink, J.-L., 2017a. Genomic Prediction in a Large African Maize Population. Crop Sci. 57, 2361. https://doi.org/10.2135/cropsci2016.08.0715

Elshire, R.J., Glaubitz, J.C., Sun, Q., Poland, J.A., Kawamoto, K., Buckler, E.S., Mitchell, S.E., 2011. A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. PLOS ONE 6, e19379. https://doi.org/10.1371/journal.pone.0019379

Endelman, J.B., 2011. Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. Plant Genome J. 4, 250. https://doi.org/10.3835/plantgenome2011.08.0024

Fong, Y.C., Khin, A.A., Lim, C.S., 2018. Conceptual Review and the Production, Consumption and Price Models of the Natural Rubber Industry in Selected ASEAN Countries and World Market. Asian J. Econ. Model. 6, 403–418. https://doi.org/10.18488/journal.8.2018.64.403.418

Ghimiray, M., Vernooy, R., 2017. The importance and challenges of crop germplasm interdependence: the case of Bhutan. Food Secur. 9, 301–310. https://doi.org/10.1007/s12571-017-0647-5

Goddard, M., 2009. Genomic selection: prediction of accuracy and maximisation of long term response. Genetica 136, 245–257. https://doi.org/10.1007/s10709-008-9308-0

Gonçalves, P. de S., Aguiar, A.T. da E., Costa, R.B. da, Gonçalves, E.C.P., Scaloppi Júnior, E.J., Branco, R.B.F., 2009. Genetic variation and realized genetic gain from rubber tree improvement. Sci. Agric. 66, 44–51. https://doi.org/10.1590/S0103-90162009000100006

Gonçalves, P. de S., Bortoletto, N., Cardinal, Á.B.B., Gouvêa, L.R.L., Costa, R.B. da, Moraes, M.L.T. de, 2005. Age-age correlation for early selection of rubber tree genotypes in São Paulo State, Brazil. Genet. Mol. Biol. 28, 758–764. https://doi.org/10.1590/S1415-47572005000500018

Gonçalves, P. de S., Martins, A.L.M., Bortoletto, N., Sáes, L.A., 2004. Selection and genetic gains for juvenile traits in progenies of Hevea in São Paulo State, Brazil. https://doi.org/10.1590/S1415-47572004000200014

Gonçalves, P. de S., Scaloppi Júnior, E.J., Martins, M.A., Moreno, R.M.B., Branco, R.B.F., Gonçalves, E.C.P., 2011. Assessment of growth and yield performance of rubber tree clones of the IAC 500 series. Pesqui. Agropecuária Bras. 46, 1643–1649. https://doi.org/10.1590/S0100-204X2011001200009

Gonçalves, P. de S., Silva, M. de A., Aguiar, A.T. da E., Martins, M.A., Junior, E.J.S., Gouvêa, L.R.L., 2007. Perfomance de novos clones de Hevea da série IAC 400. Sci. Agric. 64, 241–248. https://doi.org/10.1590/S0103-90162007000300005

Gonçalves, P. de S., Silva, M. de A., Gouvêa, L.R.L., Scaloppi Junior, E.J., 2006. Genetic variability for girth growth and rubber yield in Hevea brasiliensis. Sci. Agric. 63, 246–254. https://doi.org/10.1590/S0103-90162006000300006

Goonetilleke, S.N., March, T.J., Wirthensohn, M.G., Arús, P., Walker, A.R., Mather, D.E., 2017. Genotyping by Sequencing in Almond: SNP Discovery, Linkage Mapping, and Marker Design. G3 GenesGenomesGenetics 8, 161–172. https://doi.org/10.1534/g3.117.300376

Gorjanc, G., Battagin, M., Dumasy, J.-F., Antolín, R., Gaynor, C.R., Hickey, J.M., 2017. Prospects for Cost-Effective Genomic Selection via Accurate Within-Family Imputation. https://doi.org/10.2135/cropsci2016.06.0526

Grattapaglia, D., Resende, M.D.V., 2011. Genomic selection in forest tree breeding. Tree Genet. Genomes 7, 241–255. https://doi.org/10.1007/s11295-010-0328-4

Grinberg, N.F., Lovatt, Alan, Hegarty, M., Lovatt, Andi, Skøt, K.P., Kelly, R., Blackmore, T., Thorogood, D., King, R.D., Armstead, I., Powell, W., Skøt, L., 2016. Implementation of

Genomic Prediction in Lolium perenne (L.) Breeding Populations. Front. Plant Sci. 7. https://doi.org/10.3389/fpls.2016.00133

Habier, D., Fernando, R.L., Kizilkaya, K., Garrick, D.J., 2011. Extension of the bayesian alphabet for genomic selection. BMC Bioinformatics 12, 186. https://doi.org/10.1186/1471-2105-12-186

Hayes, B.J., Visscher, P.M., Goddard, M.E., 2009. Increased accuracy of artificial selection by using the realized relationship matrix. Genet. Res. 91, 47–60. https://doi.org/10.1017/S0016672308009981

He, J., Zhao, X., Laroche, A., Lu, Z.-X., Liu, H., Li, Z., 2014. Genotyping-by-sequencing (GBS), An ultimate marker-assisted selection (MAS) tool to accelerate plant breeding. https://doi.org/10.3389/fpls.2014.00484

He, S., Zhao, Y., Mette, M.F., Bothe, R., Ebmeyer, E., Sharbel, T.F., Reif, J.C., Jiang, Y., 2015. Prospects and limits of marker imputation in quantitative genetic studies in European elite wheat (Triticum aestivum L.). BMC Genomics 16. https://doi.org/10.1186/s12864-015-1366-y

Heff, E.L., Jannink, J.-L., Iwata, H., Souza, E.J., Sorrells, M.E., 2011. Genomic Selection Accuracy for Grain Quality Traits in Biparental Wheat Populations. https://doi.org/10.2135/cropsci2011.05.0253

Heffner, E.L., Jannink, J.-L., Iwata, H., Souza, E., Sorrells, M.E., 2011. Genomic Selection Accuracy for Grain Quality Traits in Biparental Wheat Populations. Crop Sci. 51, 2597. https://doi.org/10.2135/cropsci2011.05.0253

Heffner, E.L., Lorenz, A.J., Jannink, J.-L., Sorrells, M.E., 2010. Plant Breeding with Genomic Selection: Gain per Unit Time and Cost. Crop Sci. 50, 1681–1690. https://doi.org/10.2135/cropsci2009.11.0662

Heslot, N., Jannink, J.-L., Sorrells, M.E., 2015. Perspectives for Genomic Selection Applications and Research in Plants. Crop Sci. 55, 1. https://doi.org/10.2135/cropsci2014.03.0249

Hickey, J.M., Dreisigacker, S., Crossa, J., Hearne, S., Babu, R.S., Prasanna, B.M., Grondona, M.O., Zambelli, A., Windhausen, V.S., Mathews, K.E., Gorjanc, G., 2014. Evaluation of genomic selection training population designs and genotyping strategies in plant breeding programs using simulation. https://doi.org/10.2135/cropsci2013.03.0195

Howard, R., Carriquiry, A.L., Beavis, W.D., 2014. Parametric and Nonparametric Statistical Methods for Genomic Selection of Traits with Additive and Epistatic Genetic Architectures. G3amp58 GenesGenomesGenetics 4, 1027–1046. https://doi.org/10.1534/g3.114.010298

Howie, B.N., Donnelly, P., Marchini, J., 2009. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. 5, e1000529. https://doi.org/10.1371/journal.pgen.1000529

Iwata, H., Hayashi, T., Terakami, S., Takada, N., Sawamura, Y., Yamamoto, T., 2013. Potential assessment of genome-wide association study and genomic selection in Japanese pear Pyrus pyrifolia. Breed. Sci. 63, 125–140. https://doi.org/10.1270/jsbbs.63.125

Jahufer, M.Z.Z., Ford, J.L., Woodfield, D.R.W., Barrett, B.A., 2016. Genotypic evaluation of introduced white clover (Trifolium repens L.) germplasm in New Zealand. Crop Pasture Sci. 67, 897–906. https://doi.org/10.1071/CP16149

Jiang, Z., Wang, H., Michal, J.J., Zhou, X., Liu, B., Woods, L.C.S., Fuchs, R.A., 2016. Genome Wide Sampling Sequencing for SNP Genotyping: Methods, Challenges and Future Development. Int. J. Biol. Sci. 12, 100–108. https://doi.org/10.7150/ijbs.13498

Jiménez-Mena, B., Hospital, F., Bataillon, T., 2016. Heterogeneity in effective population size and its implications in conservation genetics and animal breeding. Conserv. Genet. Resour. 8, 35–41. https://doi.org/10.1007/s12686-015-0508-5

Jinagool, W., Rattanawong, R., Sangsing, K., Barigah, T.S., Gay, F., Cochard, H., Kasemsap, P., Herbette, S., 2015. Clonal variability for vulnerability to cavitation and other drought-related traits in Hevea brasiliensis Müll. Arg. J. Plant Hydraul. 2, 001. https://doi.org/10.20870/jph.2015.e001

Kagale, S., Koh, C., Clarke, W.E., Bollina, V., Parkin, I.A.P., Sharpe, A.G., 2016. Analysis of Genotyping-by-Sequencing (GBS) Data, in: Edwards, D. (Ed.), Plant Bioinformatics: Methods and Protocols, Methods in Molecular Biology. Springer New York, New York, NY, pp. 269–284. https://doi.org/10.1007/978-1-4939-3167-5_15

Kwong, Q.B., Teh, C.K., Ong, A.L., Chew, F.T., Mayes, S., Kulaveerasingam, H., Tammi, M., Yeoh, S.H., Appleton, D.R., Harikrishna, J.A., 2017. Evaluation of methods and marker Systems in Genomic Selection of oil palm (Elaeis guineensis Jacq.). BMC Genet. 18, 107. https://doi.org/10.1186/s12863-017-0576-5

Lau, N.-S., Makita, Y., Kawashima, M., Taylor, T.D., Kondo, S., Othman, A.S., Shu-Chien, A.C., Matsui, M., 2016. The rubber tree genome shows expansion of gene family associated with rubber biosynthesis. Sci. Rep. 6, 28594. https://doi.org/10.1038/srep28594

Lawal, B., 2003. Categorical data analysis with SAS and SPSS applications. Lawrence Erlbaum Associates, Mahwah, N.J.

Le Guen, V., Guen, V.L., Garcia, D., Doaré, F., Mattos, C.R.R., Condina, V., Couturier, C., Chambon, A., Weber, C., Espéout, S., Seguin, M., n.d. A rubber tree's durable resistance to Microcyclus ulei is conferred by a qualitative gene and a major quantitative resistance factor. Tree Genet. Amp Genomes 7, 877–889.

Leitch, A.R., Lim, K.Y., Leitch, I.J., O'Neill, M., Chye, M., Low, F., 1998. Molecular cytogenetic studies in rubber, Hevea brasiliensis Muell. Arg. (Euphorbiaceae). Genome 41, 464–467. https://doi.org/10.1139/g98-012

Lenz, P.R.N., Beaulieu, J., Mansfield, S.D., Clément, S., Desponts, M., Bousquet, J., 2017. Factors affecting the accuracy of genomic selection for growth and wood quality traits in an advanced-breeding population of black spruce (Picea mariana). BMC Genomics 18. https://doi.org/10.1186/s12864-017-3715-5

Lespinasse, D., Rodier-Goud, M., Grivet, L., Leconte, A., Legnate, H., Seguin, M., 2000. A saturated genetic linkage map of rubber tree (Hevea spp.) based on RFLP, AFLP,

microsatellite, and isozyme markers. Theor. Appl. Genet. 100, 127–138. https://doi.org/10.1007/s001220050018

Li, D., Deng, Z., Chen, C., Xia, Z., Wu, M., He, P., Chen, S., 2010. Identification and characterization of genes associated with tapping panel dryness from Hevea brasiliensis latex using suppression subtractive hybridization. BMC Plant Biol. 10, 140. https://doi.org/10.1186/1471-2229-10-140

Li, H., Rasheed, A., Hickey, L.T., He, Z., 2018. Fast-Forwarding Genetic Gain. Trends Plant Sci. 23, 184–186. https://doi.org/10.1016/j.tplants.2018.01.007

Li, Y., Dungey, H.S., 2018. Expected benefit of genomic selection over forward selection in conifer breeding and deployment. PLoS ONE 13. https://doi.org/10.1371/journal.pone.0208232

Lian, L., Jacobson, A., Zhong, S., Bernardo, R., 2014. Genomewide Prediction Accuracy within 969 Maize Biparental Populations. Crop Sci. 54, 1514. https://doi.org/10.2135/cropsci2013.12.0856

Lin, Z., Hayes, B.J., Daetwyler, H.D., 2014. Genomic selection in crops, trees and forages: a review. Crop Pasture Sci. 65, 1177–1191. https://doi.org/10.1071/CP13363

Liu, H., Zhou, H., Wu, Y., Li, X., Zhao, J., Zuo, T., Zhang, X., Zhang, Y., Liu, S., Shen, Y., Lin, H., Zhang, Z., Huang, K., Lübberstedt, T., Pan, G., 2015. The Impact of Genetic Relationship and Linkage Disequilibrium on Genomic Selection. PLOS ONE 10, e0132379. https://doi.org/10.1371/journal.pone.0132379

Liu, J.-P., Zhuang, Y.-F., Guo, X.-L., Li, Y.-J., 2016. Molecular mechanism of ethylene stimulation of latex yield in rubber tree (Hevea brasiliensis) revealed by de novo sequencing and transcriptome analysis. BMC Genomics 17. https://doi.org/10.1186/s12864-016-2587-4

Lorenz, A.J., Chao, S., Asoro, F.G., Heffner, E.L., Hayashi, T., Iwata, H., Smith, K.P., Sorrells, M.E., Jannink, J.-L., 2011. Genomic Selection in Plant Breeding: Knowledge and

Prospects, in: Donald L. Sparks (Ed.), Advances in Agronomy. Academic Press, pp. 77–123.

Luke, L.P., Sathik, M.B.M., Thomas, M., Kuruvilla, L., Sumesh, K.V., Annamalainathan, K., 2015. Quantitative expression analysis of drought responsive genes in clones of Hevea with varying levels of drought tolerance. Physiol. Mol. Biol. Plants 21, 179. https://doi.org/10.1007/s12298-015-0288-0

Ly, D., Hamblin, M., Rabbi, I., Melaku, G., Bakare, M., Gauch, H.G., Okechukwu, R., Dixon, A.G.O., Kulakow, P., Jannink, J.-L., 2013a. Relatedness and Genotype × Environment Interaction Affect Prediction Accuracies in Genomic Selection: A Study in Cassava. Crop Sci. 53, 1312. https://doi.org/10.2135/cropsci2012.11.0653

Ma, P., Brøndum, R.F., Zhang, Q., Lund, M.S., Su, G., 2013. Comparison of different methods for imputing genome-wide marker genotypes in Swedish and Finnish Red Cattle. J. Dairy Sci. 96, 4666–4677. https://doi.org/10.3168/jds.2012-6316

Makowsky, R., Pajewski, N.M., Klimentidis, Y.C., Vazquez, A.I., Duarte, C.W., Allison, D.B., de los Campos, G., 2011. Beyond missing heritability: prediction of complex traits. PLoS Genet. 7, e1002051. https://doi.org/10.1371/journal.pgen.1002051

Mantello, C.C., Suzuki, F.I., Souza, L.M., Gonçalves, P.S., Souza, A.P., 2012. Microsatellite marker development for the rubber tree (Hevea brasiliensis): characterization and cross-amplification in wild Hevea species. BMC Res. Notes 5, 329. https://doi.org/10.1186/1756-0500-5-329

Massman, J.M., Jung, H.-J.G., Bernardo, R., 2013. Genomewide Selection versus Marker-assisted Recurrent Selection to Improve Grain Yield and Stover-quality Traits for Cellulosic Ethanol in Maize. Crop Sci. 53, 58. https://doi.org/10.2135/cropsci2012.02.0112

McGowen, M.H., Vaillancourt, R.E., Pilbeam, D.J., Potts, B.M., 2010. Sources of variation in self-incompatibility in the Australian forest tree, Eucalyptus globulus. Ann. Bot. 105, 737–745. https://doi.org/10.1093/aob/mcq036

McKinnon Edwards, S., Buntjer, J.B., Jackson, R., Bentley, A.R., Lage, J., Byrne, E., Burt, C., Jack, P., Berry, S., Flatman, E., Poupard, B., Smith, S., Hayes, C., Gaynor, R., Gorjanc, G., Howell, P., Ober, E., Mackay, I.J., Hickey, J.M., 2018. The effects of training population design on genomic prediction accuracy in wheat (preprint). Genetics. https://doi.org/10.1101/443267

Mei, B., Wang, Z., 2016. An efficient method to handle the 'large p, small n' problem for genomewide association studies using Haseman–Elston regression. J. Genet. 95, 847–852. https://doi.org/10.1007/s12041-016-0705-3

Meuwissen, T.H., 2009. Accuracy of breeding values of "unrelated" individuals predicted by dense SNP genotyping. Genet. Sel. Evol. 41, 35. https://doi.org/10.1186/1297-9686-41-35

Meuwissen, T.H., Hayes, B.J., Goddard, M.E., 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157, 1819–1829.

Michel, S., Ametz, C., Gungor, H., Akgöl, B., Epure, D., Grausgruber, H., Löschenberger, F., Buerstmayr, H., 2017. Genomic assisted selection for enhancing line breeding: merging genomic and phenotypic selection in winter wheat breeding programs with preliminary yield trials. Theor. Appl. Genet. 130, 363–376. https://doi.org/10.1007/s00122-016-2818-8

Min, S., Waibel, H., Cadisch, G., Langenberger, G., Bai, J., Huang, J., 2017. The Economics of Smallholder Rubber Farming in a Mountainous Region of Southwest China: Elevation, Ethnicity, and Risk. Mt. Res. Dev. 37, 281–293. https://doi.org/10.1659/MRD-JOURNAL-D-16-00088.1

Montoro, P., Wu, S., Favreau, B., Herlinawati, E., Labrune, C., Martin-Magniette, M.-L., Pointet, S., Rio, M., Leclercq, J., Ismawanto, S., Kuswanhadi, 2018. Transcriptome analysis in Hevea brasiliensis latex revealed changes in hormone signalling pathways during ethephon stimulation and consequent Tapping Panel Dryness. Sci. Rep. 8, 8483. https://doi.org/10.1038/s41598-018-26854-y

Moraes, L.A.C., Moreira, A., Cordeiro, E.R., Moraes, V.H. de F., 2012. Translocation of cyanogenic glycosides in rubber tree crown clones resistant to South American leaf blight. Pesqui. Agropecuária Bras. 47, 906–912. https://doi.org/10.1590/S0100-204X2012000700005

Morgante, F., Huang, W., Maltecca, C., Mackay, T.F.C., 2018. Effect of genetic architecture on the prediction accuracy of quantitative traits in samples of unrelated individuals. Heredity 120, 500. https://doi.org/10.1038/s41437-017-0043-0

Muranty, H., Troggio, M., Sadok, I.B., Rifaï, M.A., Auwerkerken, A., Banchi, E., Velasco, R., Stevanato, P., van de Weg, W.E., Di Guardo, M., Kumar, S., Laurens, F., Bink, M.C.A.M., 2015. Accuracy and responses of genomic selection on key traits in apple breeding. Hortic. Res. 2, 15060. https://doi.org/10.1038/hortres.2015.60

Nielsen, N.H., Jahoor, A., Jensen, J.D., Orabi, J., Cericola, F., Edriss, V., Jensen, J., 2016. Genomic Prediction of Seed Quality Traits Using Advanced Barley Breeding Lines. PLoS ONE 11. https://doi.org/10.1371/journal.pone.0164494

Norman, A., Taylor, J., Edwards, J., Kuchel, H., 2018. Optimising Genomic Selection in Wheat: Effect of Marker Density, Population Size and Population Structure on Prediction Accuracy. G3 Bethesda Md 8, 2889–2899. https://doi.org/10.1534/g3.118.200311

Peterson, G., Dong, Y., Horbach, C., Fu, Y.-B., 2014. Genotyping-By-Sequencing for Plant Genetic Diversity Analysis: A Lab Guide for SNP Genotyping. Diversity 6, 665–680. https://doi.org/10.3390/d6040665

Pethin, D., Nakkanong, K., Nualsri, C., Pethin, D., Nakkanong, K., Nualsri, C., 2015. Performance and genetic assessment of rubber tree clones in Southern Thailand. Sci. Agric. 72, 306–313. https://doi.org/10.1590/0103-9016-2014-0354

Poets, A.M., Mohammadi, M., Seth, K., Wang, H., Kono, T.J.Y., Fang, Z., Muehlbauer, G.J., Smith, K.P., Morrell, P.L., 2015. The Effects of Both Recent and Long-Term Selection and Genetic Drift Are Readily Evident in North American Barley Breeding Populations. G3 GenesGenomesGenetics 6, 609–622. https://doi.org/10.1534/g3.115.024349

Pootakham, W., Ruang-Areerate, P., Jomchai, N., Sonthirod, C., Sangsrakru, D., Yoocha, T., Theerawattanasuk, K., Nirapathpongporn, K., Romruensukharom, P., Tragoonrung, S., Tangphatsornruang, S., 2015. Construction of a high-density integrated genetic linkage map of rubber tree (Hevea brasiliensis) using genotyping-by-sequencing (GBS). Front. Plant Sci. 6. https://doi.org/10.3389/fpls.2015.00367

Pszczola, M., Strabel, T., Mulder, H.A., Calus, M.P.L., 2012. Reliability of direct genomic values for animals with different relationships within and to the reference population. J. Dairy Sci. 95, 389–400. https://doi.org/10.3168/jds.2011-4338

R Solberg, T., Sonesson, A., Woolliams, J., H E Meuwissen, T., 2008. Genomic Selection Using Different Marker Types and Density. J. Anim. Sci. 86, 2447–54. https://doi.org/10.2527/jas.2007-0010

Rahman, A.Y.A., Usharraj, A.O., Misra, B.B., Thottathil, G.P., Jayasekaran, K., Feng, Y., Hou, S., Ong, S.Y., Ng, F.L., Lee, L.S., Tan, H.S., Sakaff, M.K.L.M., Teh, B.S., Khoo, B.F., Badai, S.S., Aziz, N.A., Yuryev, A., Knudsen, B., Dionne-Laporte, A., Mchunu, N.P., Yu, Q., Langston, B.J., Freitas, T.A.K., Young, A.G., Chen, R., Wang, L., Najimudin, N., Saito, J.A., Alam, M., 2013. Draft genome sequence of the rubber tree Hevea brasiliensis. BMC Genomics 14, 75. https://doi.org/10.1186/1471-2164-14-75

Rasheed, A., Hao, Y., Xia, X., Khan, A., Xu, Y., Varshney, R.K., He, Z., 2017. Crop Breeding Chips and Genotyping Platforms: Progress, Challenges, and Perspectives. Mol. Plant 10, 1047–1064. https://doi.org/10.1016/j.molp.2017.06.008

Ratcliffe, B., El-Dien, O.G., Klápště, J., Porth, I., Chen, C., Jaquish, B., El-Kassaby, Y.A., 2015. A comparison of genomic selection models across time in interior spruce (Picea engelmannii × glauca) using unordered SNP imputation methods. Heredity 115, 547–555. https://doi.org/10.1038/hdy.2015.57

Resende, M.D.V., Resende, M.F.R., Sansaloni, C.P., Petroli, C.D., Missiaggia, A.A., Aguiar, A.M., Abad, J.M., Takahashi, E.K., Rosado, A.M., Faria, D.A., Pappas, G.J., Kilian, A., Grattapaglia, D., 2012. Genomic selection for growth and wood quality in Eucalyptus: capturing the missing heritability and accelerating breeding for complex traits in forest

trees. New Phytol. 116–128. https://doi.org/10.1111/j.1469-8137.2011.04038.x@10.1002/(ISSN)1469-8137(CAT)FeatureIssues(VI)Bioenergytrees

Resende, M.F.R., Muñoz, P., Resende, M.D.V., Garrick, D.J., Fernando, R.L., Davis, J.M., Jokela, E.J., Martin, T.A., Peter, G.F., Kirst, M., 2012. Accuracy of Genomic Selection Methods in a Standard Data Set of Loblolly Pine (Pinus taeda L.). Genetics 190, 1503–1510. https://doi.org/10.1534/genetics.111.137026

Riedelsheimer, C., Endelman, J.B., Stange, M., Sorrells, M.E., Jannink, J.-L., Melchinger, A.E., 2013. Genomic predictability of interconnected biparental maize populations. Genetics 194, 493–503. https://doi.org/10.1534/genetics.113.150227

Rivano, F., Mattos, C.R.R., Cardoso, S.E.A., Martinez, M., Cevallos, V., Le Guen, V., Garcia, D., 2013. Breeding Hevea brasiliensis for yield, growth and SALB resistance for high disease environments. Ind. Crops Prod. 44, 659–670. https://doi.org/10.1016/j.indcrop.2012.09.005

Robertsen, C.D., Hjortshøj, R.L., Janss, L.L., 2019. Genomic Selection in Cereal Breeding. Agronomy 9, 95. https://doi.org/10.3390/agronomy9020095

Ronald, P., 2011. Plant Genetics, Sustainable Agriculture and Global Food Security. Genetics 188, 11–20. https://doi.org/10.1534/genetics.111.128553

Rose, K., Steinbüchel, A., 2005. Biodegradation of Natural Rubber and Related Compounds: Recent Insights into a Hardly Understood Catabolic Capability of Microorganisms. Appl. Environ. Microbiol. 71, 2803–2812. https://doi.org/10.1128/AEM.71.6.2803-2812.2005

Rutkoski, J.E., Poland, J., Jannink, J.-L., Sorrells, M.E., 2013. Imputation of Unordered Markers and the Impact on Genomic Selection Accuracy. G3 GenesGenomesGenetics 3, 427–439. https://doi.org/10.1534/g3.112.005363

Sainoi, T., Sdoodee, S., Lacote, R., Gohet, E., 2017. Low frequency tapping systems applied to young-tapped trees of Hevea brasiliensis (Willd. ex A. Juss.) Müll. Arg. in Southern Thailand. Agric. Nat. Resour. 51, 268–272. https://doi.org/10.1016/j.anres.2017.03.001

Sakdapipanich, J.T., Rojruthai, P., 2012. Molecular Structure of Natural Rubber and Its Characteristics Based on Recent Evidence. Biotechnol. - Mol. Stud. Nov. Appl. Improv. Qual. Hum. Life. https://doi.org/10.5772/29820

Sarinelli, J.M., Murphy, J.P., Tyagi, P., Holland, J.B., Johnson, J.W., Mergoum, M., Mason, R.E., Babar, A., Harrison, S., Sutton, R., Griffey, C.A., Brown-Guedira, G., 2019. Training population selection and use of fixed effects to optimize genomic predictions in a historical USA winter wheat panel. Theor. Appl. Genet. 132, 1247–1261. https://doi.org/10.1007/s00122-019-03276-6

Schopp, P., Müller, D., Wientjes, Y.C.J., Melchinger, A.E., 2017. Genomic Prediction Within and Across Biparental Families: Means and Variances of Prediction Accuracy and Usefulness of Deterministic Equations. G3 GenesGenomesGenetics 7, 3571–3586. https://doi.org/10.1534/g3.117.300076

Shamshad, M., Sharma, A., 2018. The Usage of Genomic Selection Strategy in Plant Breeding. Gener. Plant Breed. https://doi.org/10.5772/intechopen.76247

Sniezko, R., Smith, J., Liu, J.-J., Hamelin, R., 2014. Genetic Resistance to Fusiform Rust in Southern Pines and White Pine Blister Rust in White Pines—A Contrasting Tale of Two Rust Pathosystems—Current Status and Future Prospects. Forests 5, 2050–2083. https://doi.org/10.3390/f5092050

Sonah, H., O'Donoughue, L., Cober, E., Rajcan, I., Belzile, F., 2015. Identification of loci governing eight agronomic traits using a GBS-GWAS approach and validation by QTL mapping in soya bean. Plant Biotechnol. J. 13, 211–221. https://doi.org/10.1111/pbi.12249

Sorrells, M.E., 2015. Genomic Selection in Plants: Empirical Results and Implications for Wheat Breeding, in: Ogihara, Y., Takumi, S., Handa, H. (Eds.), Advances in Wheat Genetics: From Genome to Field. Springer Japan, pp. 401–409.

Sousa, T.V., Caixeta, E.T., Alkimim, E.R., Oliveira, A.C.B., Pereira, A.A., Sakiyama, N.S., Zambolim, L., Resende, M.D.V., 2019. Early Selection Enabled by the Implementation

of Genomic Selection in Coffea arabica Breeding. Front. Plant Sci. 9, 1934. https://doi.org/10.3389/fpls.2018.01934

Souza, A., Regina Lima Gouvêa, L., Oliveira, A., Silva, G., Gonçalves, P., 2017. Associations among rubber yield and secondary traits in juvenile rubber trees progeny. Euphytica 213. https://doi.org/10.1007/s10681-016-1804-1

Souza, L.M., Gazaffi, R., Mantello, C.C., Silva, C.C., Garcia, D., Guen, V.L., Cardoso, S.E.A., Garcia, A.A.F., Souza, A.P., 2013. QTL Mapping of Growth-Related Traits in a Full-Sib Family of Rubber Tree (Hevea brasiliensis) Evaluated in a Sub-Tropical Climate. PLOS ONE 8, e61238. https://doi.org/10.1371/journal.pone.0061238

Stam, P., 1993. Construction of integrated genetic linkage maps by means of a new computer package: Join Map. Plant J. 3, 739–744. https://doi.org/10.1111/j.1365-313X.1993.00739.x

Stekhoven, D.J., Bühlmann, P., 2012. MissForest - nonparametric missing value imputation for mixed-type data. Bioinformatics 28, 112–118. https://doi.org/10.1093/bioinformatics/btr597

Stich, B., Van Inghelandt, D., 2018. Prospects and Potential Uses of Genomic Prediction of Key Performance Traits in Tetraploid Potato. Front. Plant Sci. 9. https://doi.org/10.3389/fpls.2018.00159

Sun, C., Wu, X.-L., Weigel, K.A., Rosa, G.J.M., Bauck, S., Woodward, B.W., Schnabel, R.D., Taylor, J.F., Gianola, D., 2012. An ensemble-based approach to imputation of moderate-density genotypes for genomic selection with application to Angus cattle. Genet. Res. 94, 133–150. https://doi.org/10.1017/S001667231200033X

Suontama, M., Klápště, J., Telfer, E., Graham, N., Stovold, T., Low, C., McKinley, R., Dungey, H., 2019. Efficiency of genomic prediction across two Eucalyptus nitens seed orchards with different selection histories. Heredity 122, 370–379. https://doi.org/10.1038/s41437-018-0119-5

Syahputri, K., Sari, R.M., Rizkya, I., Siregar, I., 2017. Identification and Waste Reduction on Rubber Industry. IOP Conf. Ser. Mater. Sci. Eng. 180, 012119. https://doi.org/10.1088/1757-899X/180/1/012119

Tanielian, A., 2018. Sustainability and Competitiveness in Thai Rubber Industries. Cph. J. Asian Stud. 36, 50. https://doi.org/10.22439/cjas.v36i1.5512

Torkamaneh, D., Laroche, J., Bastien, M., Abed, A., Belzile, F., 2017. Fast-GBS: a new pipeline for the efficient and highly accurate calling of SNPs from genotyping-by-sequencing data. BMC Bioinformatics 18, 5. https://doi.org/10.1186/s12859-016-1431-9

Umar, H.Y., Giroh, D.Y., Agbonkpolor, N.B., Mesike, C.S., 2011. An Overview of World Natural Rubber Production and Consumption: An Implication for Economic Empowerment and Poverty Alleviation in Nigeria. J. Hum. Ecol. 33, 53–59. https://doi.org/10.1080/09709274.2011.11906350

Venkatachalam, P., Thulaseedharan, A., Raghothama, K., 2009. Molecular identification and characterization of a gene associated with the onset of tapping panel dryness (TPD) syndrome in rubber tree (Hevea brasiliensis Muell.) by mRNA differential display. Mol. Biotechnol. 41, 42–52. https://doi.org/10.1007/s12033-008-9095-y

Viana, J.M.S., Piepho, H.-P., Silva, F.F. e, 2016. Quantitative genetics theory for genomic selection and efficiency of breeding value prediction in open-pollinated populations. Sci. Agric. 73, 243–251. https://doi.org/10.1590/0103-9016-2014-0383

Visscher, P.M., Medland, S.E., Ferreira, M.A., Morley, K.I., Zhu, G., Cornes, B.K., Montgomery, G.W., Martin, N.G., 2006. Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. PLoS Genet. 2, e41–e41. https://doi.org/10.1371/journal.pgen.0020041

Voss-Fels, K.P., Cooper, M., Hayes, B.J., 2018. Accelerating crop genetic gains with genomic selection. Theor. Appl. Genet. https://doi.org/10.1007/s00122-018-3270-8

Wang, Q., Yu, Y., Yuan, J., Zhang, X., Huang, H., Li, F., Xiang, J., 2017. Effects of marker density and population structure on the genomic prediction accuracy for growth trait in

Pacific white shrimp Litopenaeus vannamei. BMC Genet. 18, 45. https://doi.org/10.1186/s12863-017-0507-5

Wang, X., Xu, Y., Hu, Z., Xu, C., 2018. Genomic selection methods for crop improvement: Current status and prospects. Crop J. 6, 330–340. https://doi.org/10.1016/j.cj.2018.03.001

Warren-Thomas, E., Dolman, P.M., Edwards, D.P., 2015. Increasing Demand for Natural Rubber Necessitates a Robust Sustainability Initiative to Mitigate Impacts on Tropical Biodiversity. Conserv. Lett. 8, 230–241. https://doi.org/10.1111/conl.12170

Warren-Thomas, E.M., Edwards, D.P., Bebber, D.P., Chhang, P., Diment, A.N., Evans, T.D., Lambrick, F.H., Maxwell, J.F., Nut, M., O'Kelly, H.J., Theilade, I., Dolman, P.M., 2018. Protecting tropical forests from the rapid expansion of rubber using carbon payments. Nat. Commun. 9, 911. https://doi.org/10.1038/s41467-018-03287-9

Weng, Z., Zhang, Z., Zhang, Q., Fu, W., He, S., Ding, X., 2013. Comparison of different imputation methods from low- to high-density panels using Chinese Holstein cattle. Animal 7, 729–735. https://doi.org/10.1017/S1751731112002224

Werner, C.R., Voss-Fels, K.P., Miller, C.N., Qian, W., Hua, W., Guan, C.-Y., Snowdon, R.J., Qian, L., 2018. Effective Genomic Selection in a Narrow-Genepool Crop with Low-Density Markers: Asian Rapeseed as an Example. Plant Genome 11. https://doi.org/10.3835/plantgenome2017.09.0084

Whittaker, J.C., Thompson, R., Denham, M.C., 2000. Marker-assisted selection using ridge regression. Genet. Res. 75, 249–252.

Wickland, D.P., Battu, G., Hudson, K.A., Diers, B.W., Hudson, M.E., 2017. A comparison of genotyping-by-sequencing analysis methods on low-coverage crop datasets shows advantages of a new workflow, GB-eaSy. BMC Bioinformatics 18, 586. https://doi.org/10.1186/s12859-017-2000-6

Wientjes, Y.C.J., Veerkamp, R.F., Calus, M.P.L., 2013. The Effect of Linkage Disequilibrium and Family Relationships on the Reliability of Genomic Prediction. Genetics 193, 621–631. https://doi.org/10.1534/genetics.112.146290

Wu, M., McIntosh, J., Liu, J., 2016. Current prevalence rate of latex allergy: Why it remains a problem? J. Occup. Health 58, 138–144.

Yamamoto, E., Matsunaga, H., Onogi, A., Ohyama, A., Miyatake, K., Yamaguchi, H., Nunome, T., Iwata, H., Fukuoka, H., 2017. Efficiency of genomic selection for breeding population design and phenotype prediction in tomato. Heredity 118, 202–209. https://doi.org/10.1038/hdy.2016.84

Yang, S., Fresnedo-Ramírez, J., Wang, M., Cote, L., Schweitzer, P., Barba, P., Takacs, E.M., Clark, M., Luby, J., Manns, D.C., Sacks, G., Mansfield, A.K., Londo, J., Fennell, A., Gadoury, D., Reisch, B., Cadle-Davidson, L., Sun, Q., 2016. A next-generation marker genotyping platform (AmpSeq) in heterozygous crops: a case study for marker-assisted selection in grapevine. Hortic. Res. 3, 16002. https://doi.org/10.1038/hortres.2016.2

Yeang, H.-Y., 2007. Synchronous flowering of the rubber tree (Hevea brasiliensis) induced by high solar radiation intensity. New Phytol. 175, 283–289. https://doi.org/10.1111/j.1469-8137.2007.02089.x

Zhang, A., Wang, H., Beyene, Y., Semagn, K., Liu, Y., Cao, S., Cui, Z., Ruan, Y., Burgueño, J., San Vicente, F., Olsen, M., Prasanna, B.M., Crossa, J., Yu, H., Zhang, X., 2017. Effect of Trait Heritability, Training Population Size and Marker Density on Genomic Prediction Accuracy Estimation in 22 bi-parental Tropical Maize Populations. Front. Plant Sci. 8, 1916. https://doi.org/10.3389/fpls.2017.01916

Zhang, H., Yin, L., Wang, M., Yuan, X., Liu, X., 2019. Factors Affecting the Accuracy of Genomic Selection for Agricultural Economic Traits in Maize, Cattle, and Pig Populations. Front. Genet. 10, 189. https://doi.org/10.3389/fgene.2019.00189

Zhao, J., Han, D., Shi, K., Wang, L., Gao, J., Yang, R., 2018. Influence of epistatic segregation distortion loci on genetic marker linkages in Japanese flounder. Genomics 110, 59–66. https://doi.org/10.1016/j.ygeno.2017.08.006

Zhao, Y., Mette, M.F., Gowda, M., Longin, C.F.H., Reif, J.C., 2014. Bridging the gap between marker-assisted and genomic selection of heading time and plant height in hybrid wheat. Heredity 112, 638–645. https://doi.org/10.1038/hdy.2014.1

Zhao, Y., Zeng, J., Fernando, R., Reif, J.C., 2013. Genomic Prediction of Hybrid Wheat Performance. Crop Sci. 53, 802–810. https://doi.org/10.2135/cropsci2012.08.0463

APPENDICES

Appendix 1: Genetic linkage map of progeny of the cross PB260 × RRIM600

Locus	Segregation	Group	Position (cM)	Marker
L843	<hkxhk>	g01	0.00	scaffold0168:1086506_A/G
L1356	<hkxhk>	g01	2.02	scaffold0326:201227_C/G
L3162	<hkxhk>	g01	4.02	scaffold1700:13616_C/T
g1a134	<efxeg>	g01	6.78	g1a134
g1a274	<lmxll>	g01	7.23	g1a274
g1t319	<efxeg>	g01	8.53	g1t319
g1A2705	<hkxhk>	g01	10.09	g1A2705
g1A2497	<abxcd>	g01	10.59	g1A2497
g1A2746	<abxcd>	g01	10.69	g1A2746
L2522	<hkxhk>	g01	43.55	scaffold0872:408342_T/C
L2322	<hkxhk>	g01	44.42	scaffold0746:324277_C/T
L2321	<hkxhk>	g01	46.50	scaffold0746:116707_G/T
L946	<hkxhk>	g01	48.23	scaffold0195:1001662_G/T
L941	<hkxhk>	g01	49.30	scaffold0195:463803_T/C
L940	<hkxhk>	g01	51.07	scaffold0195:381781_T/C
L645	<hkxhk>	g01	53.53	scaffold0113:1897644_T/C
L2901	<hkxhk>	g01	54.61	scaffold1196:208986_C/T
g1SSH033	<abxcd>	g01	105.78	g1SSH033
L886	<hkxhk>	g01	106.36	scaffold0179:1251050_C/T
L2021	<lmxll>	g01	107.09	scaffold0581:142388_G/A
L3354	<lmxll>	g01	108.17	scaffold2730:8125_A/G
L64	<hkxhk>	g01	108.96	scaffold0008:4159882_T/C
L885	<hkxhk>	g01	109.83	scaffold0179:1206342_T/C
L76	<hkxhk>	g01	110.50	scaffold0010:4112757_A/G
L542	<nnxnp>	g01	111.35	scaffold0090:1942100_C/T
L75	<hkxhk>	g01	112.38	scaffold0010:4112209_C/G
L725	<hkxhk>	g01	114.21	scaffold0135:1452303_C/A
L2412	<hkxhk>	g01	114.86	scaffold0809:18562_C/T
L3181	<hkxhk>	g01	115.97	scaffold1731:54803_T/C
L718	<nnxnp>	g01	117.34	scaffold0133:768735_G/A
L1915	<lmxll>	g01	117.78	scaffold0543:842733_A/G
L3409	<hkxhk>	g01	119.07	scaffold4381:5798_G/A
L3248	<lmxll>	g01	121.69	scaffold2013:36869_T/G
L3083	<hkxhk>	g01	123.49	scaffold1532:94876_T/A
L8	<hkxhk>	g01	126.63	scaffold0001:5858685_C/T

L3187	<nnxnp>	g01	130.46	scaffold1754:4923_G/A
L224	<hkxhk>	g02	0.00	scaffold0029:1513004_A/T
L1177	<hkxhk>	g02	1.92	scaffold0266:1090363_A/T
L1176	<hkxhk>	g02	2.95	scaffold0266:1035328_A/T
L223	<hkxhk>	g02	3.91	scaffold0029:1509715_C/T
L24	<lmxll>	g02	7.62	scaffold0003:5339561_C/T
L2532	<hkxhk>	g02	7.82	scaffold0878:342412_A/G
L1379	<hkxhk>	g02	10.76	scaffold0337:1216058_T/C
g2BAC12N03rev	<abxcd>	g02	12.71	g2BAC12N03rev
L95	<hkxhk>	g02	14.09	scaffold0013:26364_T/C
L98	<lmxll>	g02	16.21	scaffold0013:180318_G/A
L1096	<hkxhk>	g02	17.94	scaffold0239:1245236_T/C
L1095	<hkxhk>	g02	21.30	scaffold0239:685526_A/T
L3061	<hkxhk>	g02	35.92	scaffold1490:36216_A/C
L3048	<hkxhk>	g02	36.54	scaffold1460:54106_A/G
L356	<lmxll>	g02	36.66	scaffold0053:19006_G/A
L3046	<hkxhk>	g02	37.74	scaffold1460:50389_A/G
L3045	<hkxhk>	g02	38.77	scaffold1460:46923_G/A
L3044	<hkxhk>	g02	39.23	scaffold1460:8206_C/T
L3411	<hkxhk>	g02	40.04	scaffold4699:4294_A/C
L924	<hkxhk>	g02	40.76	scaffold0189:1467387_C/A
g2A2734	<efxeg>	g02	41.61	g2A2734
L952	<nnxnp>	g02	41.85	scaffold0196:35618_T/G
L1753	<hkxhk>	g02	42.82	scaffold0470:273465_T/C
g2A2680.L1	<efxeg>	g02	45.69	g2A2680.L1
g2T2607	<lmxll>	g02	45.70	g2T2607
L519	<lmxll>	g02	66.70	scaffold0085:1731869_T/C
L1799	<lmxll>	g02	67.77	scaffold0493:64101_A/T
L1161	<lmxll>	g02	68.16	scaffold0258:16515_T/G
L2859	<lmxll>	g02	70.15	scaffold1155:58979_A/T
L712	<hkxhk>	g02	70.88	scaffold0131:706488_G/T
L1795	<lmxll>	g02	71.34	scaffold0490:204904_T/C
L344	<lmxll>	g02	81.43	scaffold0050:390797_C/T
L94	<hkxhk>	g02	81.98	scaffold0012:3603756_T/C
L309	<lmxll>	g02	82.18	scaffold0045:1097675_A/G
L3370	<hkxhk>	g02	82.95	scaffold3076:4527_T/C
L304	<lmxll>	g02	83.37	scaffold0043:2161629_T/C
L3104	<hkxhk>	g02	85.13	scaffold1564:70940_T/A
L324	<nnxnp>	g02	85.40	scaffold0047:1741007_C/A
g2T2094	<abxcd>	g02	86.16	g2T2094
L824	<lmxll>	g02	87.65	scaffold0164:305713_A/G
L554	<hkxhk>	g02	88.90	scaffold0093:2350513_G/A
L2175	<lmxll>	g02	89.40	scaffold0654:321423_C/A

g2A2381	<efxeg>	g02	90.83	g2A2381
g2t152	<abxcd>	g02	90.98	g2t152
g2t283	<abxcd>	g02	91.76	g2t283
L2057	<lmxll>	g02	92.47	scaffold0593:250124_G/A
L1165	<hkxhk>	g02	96.83	scaffold0261:1396795_C/G
L1762	<lmxll>	g02	97.46	scaffold0474:967774_T/C
L823	<hkxhk>	g02	99.46	scaffold0164:224406_C/G
L1141	<hkxhk>	g03	0.00	scaffold0251:247346_T/C
L1142	<hkxhk>	g03	2.89	scaffold0251:394559_A/C
L1140	<hkxhk>	g03	4.49	scaffold0251:131794_A/G
L3252	<hkxhk>	g03	6.66	scaffold2028:20921_C/G
L1351	<nnxnp>	g03	7.76	scaffold0323:406028_G/A
g3A2707	<abxcd>	g03	46.33	g3A2707
L1212	<hkxhk>	g03	49.98	scaffold0280:31068_G/A
L609	<nnxnp>	g03	54.34	scaffold0104:1561871_T/C
L2400	<lmxll>	g03	55.85	scaffold0802:99062_C/G
L2645	<hkxhk>	g03	60.52	scaffold0959:245013_A/C
L1255	<hkxhk>	g03	64.07	scaffold0297:1274154_A/G
g3TAs2697	<efxeg>	g03	66.06	g3TAs2697
L688	<hkxhk>	g03	67.91	scaffold0123:1900589_A/G
L1953	<lmxll>	g03	69.41	scaffold0556:229667_T/C
L818	<hkxhk>	g03	71.04	scaffold0163:28896_C/T
g3A312	<hkxhk>	g03	71.85	g3A312
L2391	<nnxnp>	g03	71.91	scaffold0795:12323_G/A
L573	<hkxhk>	g03	73.68	scaffold0098:789896_T/C
L574	<hkxhk>	g03	74.98	scaffold0098:1034125_C/T
L2644	<lmxll>	g03	75.22	scaffold0959:145589_C/T
g3SSH031.L2	<lmxll>	g03	76.75	g3SSH031.L2
L2272	<hkxhk>	g03	77.88	scaffold0711:528714_C/T
L2671	<hkxhk>	g03	79.56	scaffold0988:71447_C/A
L3258	<hkxhk>	g03	93.22	scaffold2054:44152_G/T
L672	<hkxhk>	g03	93.77	scaffold0118:1862442_C/T
g3a403	<nnxnp>	g03	94.84	g3a403
g3SSH059	<lmxll>	g03	94.98	g3SSH059
L2333	<hkxhk>	g03	96.60	scaffold0756:435250_T/C
L1549	<lmxll>	g03	96.63	scaffold0400:431053_A/G
L2332	<hkxhk>	g03	97.89	scaffold0756:304301_T/A
L1078	<hkxhk>	g03	98.56	scaffold0235:90271_T/C
L1079	<hkxhk>	g03	99.41	scaffold0235:233569_C/G

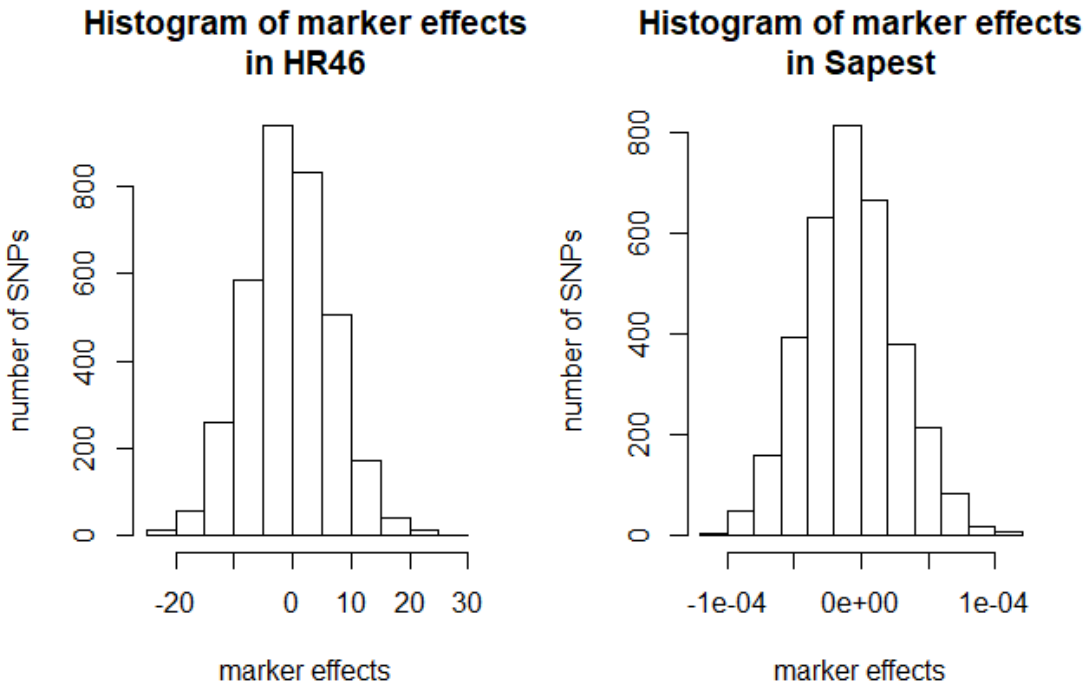
Appendix 2: Distribution of SSR markers on the Hevea genetic map of the progeny of a cross of PB260 × RRIM600

Linkage group	Number of SSR Markers	Number of SNP Markers	Total	Length in cM
LG1	13	95	108	130.5
LG2	21	105	126	131.7
LG3	15	99	114	152.7
LG4	12	59	71	137.7
LG5	20	143	163	147.9
LG6a	12	51	63	114.1
LG6b	2	20	22	42.1
LG7	18	78	96	165.9
LG8	22	84	106	148.1
LG9	23	112	135	155.4
LG10	32	139	171	181.8
LG11	13	120	133	156.5
LG12	10	86	96	111.0
LG13	15	108	123	150.3
LG14	19	124	143	134.3
LG15	18	112	130	171.9
LG16	14	70	84	135.7
LG17	16	68	84	101.4
LG18	13	96	109	132.1
Total	308	1769	2077	2600.9

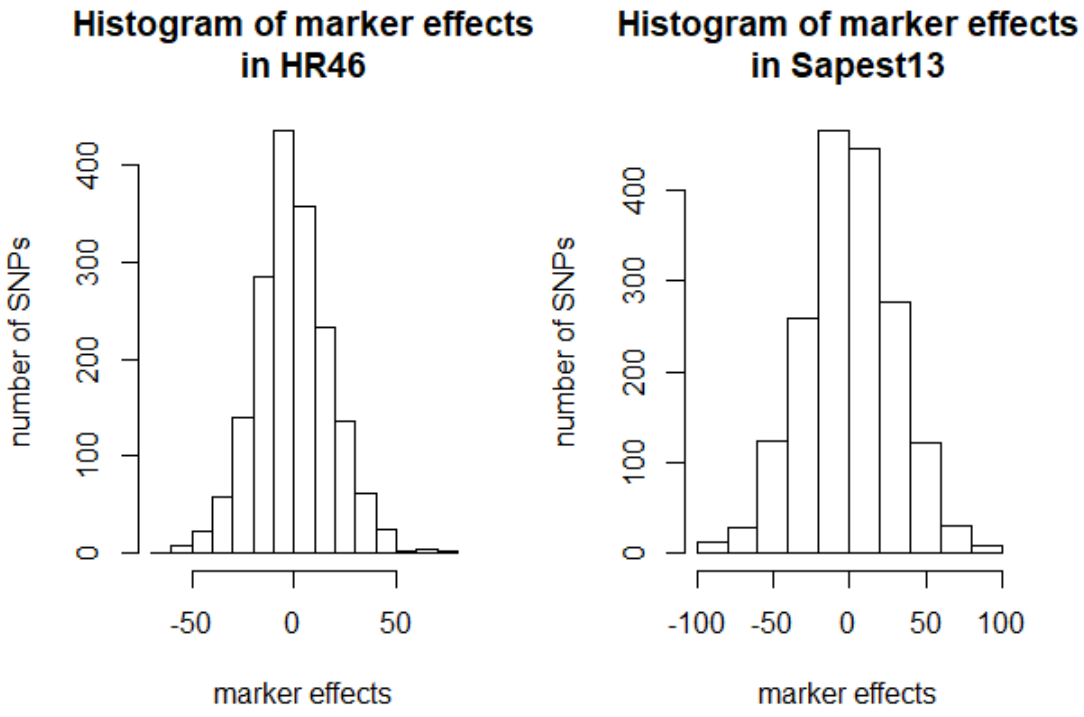
Appendix 3: **Segregation of SSR and SNP markers on the Hevea genetic map of progeny of a cross of PB260 × RRIM600**

Marker segregation type	Number of SSR markers	Number of SNP markers
<abxcd>	42	0
<efxeg>	88	0
<hkxhk>	39	1339
<lmxll>	89	269
<nnxnp>	50	161
Total	308	1769

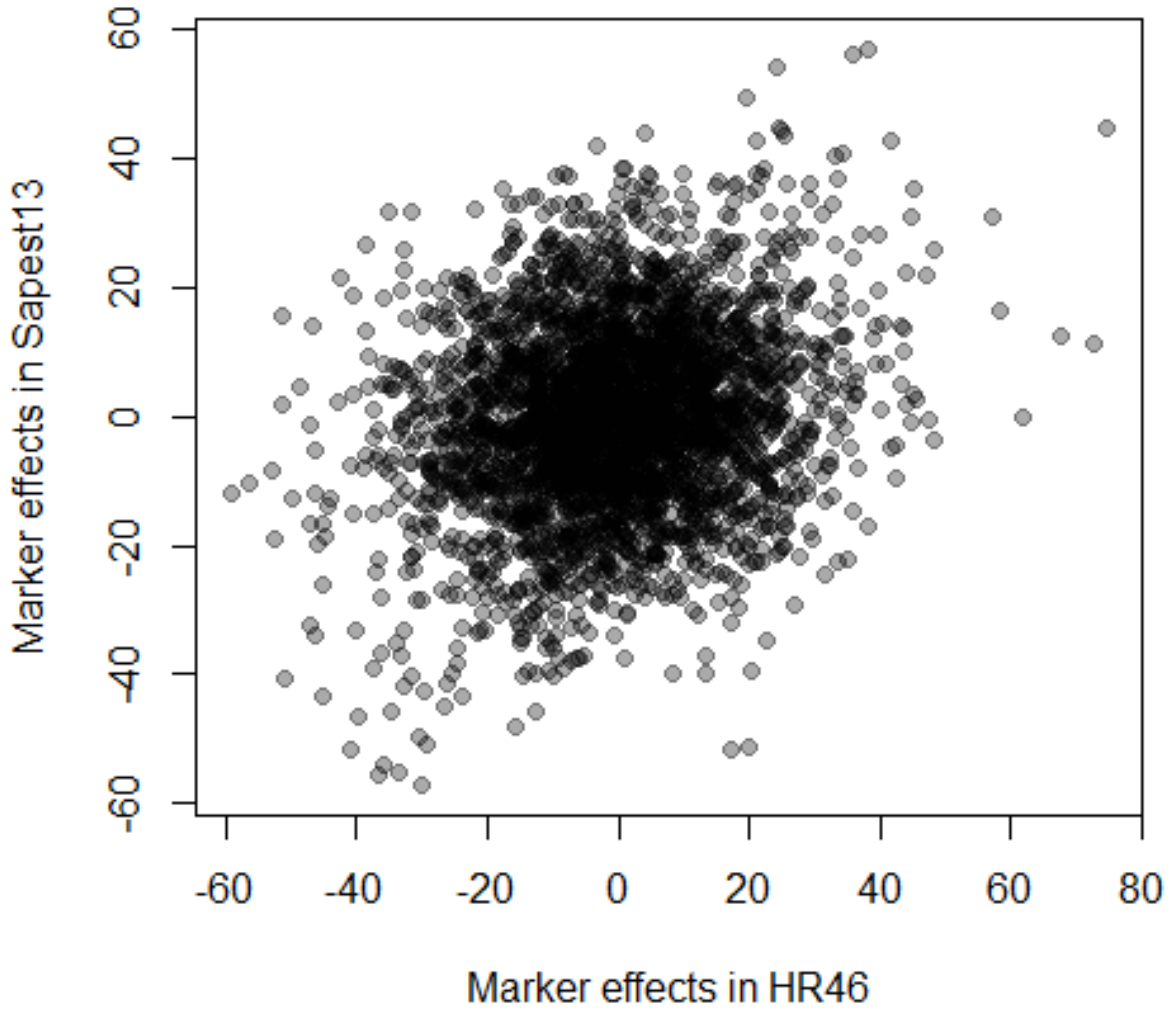
Appendix 4: **Distribution of marker effects using marker data from Random Forest imputation**



Appendix 5: **Distribution of marker effects using marker data from Beagle imputation**



Appendix 6: **Correlation between marker effects in the two sites (HR46 and Sapest13) using markers from Random Forest imputation.**



Appendix 7: Correlation between marker effects in the two sites (HR46 and Sapest13) using markers from Beagle imputation