



# Genomic and pedigree-based predictive ability for quality traits in tea (*Camellia sinensis* (L.) O. Kuntze)

Nelson Lubanga · Festo Massawe · Sean Mayes

Received: 14 April 2020 / Accepted: 16 January 2021  
© The Author(s) 2021

**Abstract** Genetic improvement of quality traits in tea (*Camellia sinensis* (L.) O. Kuntze) through conventional breeding methods has been limited, because tea quality is a difficult and expensive trait to measure. Genomic selection (GS) is suitable for predicting such complex traits, as it uses genome wide markers to estimate the genetic values of individuals. We compared the prediction accuracies of six genomic prediction models including Bayesian ridge regression (BRR), genomic best linear unbiased prediction (GBLUP), BayesA, BayesB, BayesC and reproducing kernel Hilbert spaces models incorporating the pedigree relationship namely; RKHS-pedigree, RKHS-markers and RKHS markers and pedigree (RKHS-MP) to determine the breeding values for 12 tea

quality traits. One hundred and three tea genotypes were genotyped using genotyping-by-sequencing and phenotyped using nuclear magnetic resonance spectroscopy in replicated trials. We also compared the effect of trait heritability and training population size on prediction accuracies. The traits with the highest prediction accuracies were; theogallin (0.59), epicatechin gallate (ECG) (0.56) and theobromine (0.61), while the traits with the lowest prediction accuracies were theanine (0.32) and caffeine (0.39). The performance of all the GS models were almost the same, with BRR (0.53), BayesA (0.52), GBLUP (0.50) and RKHS-MP (0.50) performing slightly better than the others. Heritability estimates were moderate to high (0.35–0.92). Prediction accuracies increased with increasing training population size and trait heritability. We conclude that the moderate to high prediction accuracies observed suggests GS is a promising

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1007/s10681-021-02774-3>.

N. Lubanga (✉)  
Roslin Institute, The University of Edinburgh, Easter  
Bush Campus, Midlothian EH25 9RG, UK  
e-mail: nelsonlubanga@gmail.com

N. Lubanga  
Unilever Tea Kenya Ltd., Nakuru Road,  
Kericho 20-20200, Kenya

N. Lubanga  
Unilever Research and Development, Colworth Science  
Park, Sharnbrook, Bedford MK441LQ, UK

N. Lubanga · F. Massawe  
The University of Nottingham Malaysia, Jalan Broga,  
43500 Semenyih, Selangor Darul Ehsan, Malaysia

S. Mayes  
School of Biosciences, The University of Nottingham,  
Sutton Bonington Campus, Loughborough,  
Leicestershire LE12 5RD, UK

approach in tea improvement and could be implemented in breeding programmes.

**Keywords** Tea breeding · Genomic selection · Tea quality

## Introduction

Tea (*Camellia sinensis* (L.) O. Kuntze) quality is an important attribute in a tea breeding programme. It is the main determinant of price at the tea auction and is measured based on the flavour and colour of the liquor (hue) along with appearance of dry tea (leaf) (Zheng et al. 2016). Flavour comprises of taste, mouthfeel and aroma (Lawless and Heymann 2010). Taste of tea is characterized by the astringency, bitterness, mellowness and slight sweetness (Lee and Chambers 2007). Mouth-feel is the heaviness, thickness and strength of tea liquor, while aroma is influenced by more than 600 volatile compounds known to be present in tea (Zheng et al. 2016). Taste, mouthfeel, colour and aroma are important tea quality traits for consumer and are key targets for selection in breeding programmes. These tea attributes originate from biochemical compounds present in fresh tea shoots such as catechins, alkaloids, amino acids and volatile compounds (Borse 2012; Chen et al. 2018a).

Genomic selection (GS) is a modern breeding approach whereby models based on genome-wide markers are used to estimate marker effects across the entire genome to produce an estimate of the genetic values (Jannink et al. 2010; Meuwissen et al. 2001). GS models attempt to capture total additive genetic variance across the entire genome to estimate GEBVs among the selection candidates based on the sum of all marker effects (Lorenz et al. 2011a). Genomic estimated breeding values (GEBVs) of the next generation of untested genotypes with only genotypic information are computed using the constructed model and these are used for selection of superior individuals without direct phenotypic evaluation (Meuwissen et al. 2001). In GS, the number of markers are usually greater than the number of phenotypic measurements of the traits of interest, hence there are more predictor variables compared to phenotypes, hence creating a “large p and small n problem” (Heffner et al. 2011). Statistical models that

have been developed to solve the problem of having large numbers of molecular markers and fewer phenotypes include ridge regression best linear unbiased predictor (rrBLUP), genomic best linear unbiased predictor (G-BLUP), the Bayesian models (BayesA, BayesB, BayesC, BayesLASSO) and machine learning models (Wang et al. 2018). RRBLUP is computationally similar to genomic BLUP (GBLUP) and it assumes that marker effects are equally shrunk and normally distributed with the same variance (Meuwissen et al. 2001). It is an infinitesimal model and assumes that all the markers have small effects and have non-zero variance. On the other hand, Bayesian models assume the markers have different amounts of variation and are more flexible while predicting traits with different genetic architectures (Habier et al. 2011). Bayesian models are therefore suited for traits that are controlled by few large-effect genes compared to RRBLUP (Beaulieu et al. 2014; Meuwissen et al. 2001). BayesA and BayesLASSO assume that all markers have a non-zero effect, and the marker variances are derived from a scaled inverted chi-square and double-exponential distributions, respectively. Both BayesB and BayesC are variable selection models since they are derived from two component mixtures with a point of mass at zero that can either be a scaled-t or a normal distributions, respectively (Habier et al. 2011). The reproducing kernel Hilbert spaces model (RKHS) is a semi-parametric approach for genomic prediction and several studies have shown its effectiveness in genomic predictions (Crossa et al. 2010; Juliana et al. 2017). It does not assume linearity and therefore also captures some non-additive effects well (Juliana et al. 2017).

GS models have successfully been developed for predicting traits for many crops (Bassi et al. 2016; Cerrudo et al. 2018; El-Dien et al. 2015; Grattapaglia et al. 2018; Juliana et al. 2017; Müller et al. 2019; Sverrisdóttir et al. 2017; Tan et al. 2017). GS can potentially reduce the length of the tea breeding cycle in tea and increase gains per unit time through early selection, with the GS model being used to carry out 1–2 rounds of selection based on genotype alone, before the need to rebuild the model due to the change in allelic frequencies caused by selection. Koech et al. (2020) applied machine learning models to estimate the prediction accuracies of black tea quality and drought tolerance traits in discovery and validation populations. However, they used a limited number of

markers (i.e. 1,421 DarTseq markers) and only machine learning models were compared. There are no reported studies of genomic selection in tea using parametric models. Additionally, there is no evidence of GS implementation in a tea breeding programme. While several studies comparing the performance of different prediction models have been reported in many crops (Grattapaglia et al. 2018; Kwong et al. 2017; Lozada et al. 2019), our objective was to compare the prediction accuracies of six genomic prediction models including Bayesian Ridge Regression (BRR), genomic best linear unbiased prediction (GBLUP), BayesA, BayesB, BayesC and reproducing kernel Hilbert spaces models incorporating the pedigree relationship namely; RKHS-pedigree (RKHS-P), RKHS-markers (RKHS-M) and RKHS markers and pedigree (RKHS-MP), to determine the breeding values for 12 tea quality traits measured in two different environments using Nuclear Magnetic Resonance (NMR). We also evaluated the effects of training population size and heritability on the accuracy of GS. Lastly, we discussed how GS can be implemented in a tea breeding programme.

## Materials and methods

### Plant materials and phenotyping

Genotypes used in this study consisted of 103 tea varieties (clones), present in the UTK breeding programme clonal field trials (CFTs) at Kericho (0°22' S and 35°17' E), which is located at 2005 meters above sea level and replicated at Jamji (0°28' S and 35°11' E), situated at 1733 meters above sea level. Three replicates of each genotype was then phenotyped at each site using nuclear magnetic resonance (NMR) spectroscopy for the 12 quality traits namely; theobromine, caffeine, theogallin, gallic acid (GA), epicatechin (EC), galocatechin gallate (GCG), epicatechin gallate (ECG), epigallocatechin gallate (EGCG), epigallocatechin (EGC), theanine, catechin (C) and galocatechin (GC) according to Le Gall et al. (2004). Analysis of variance was conducted for all the traits to estimate significant differences between the genotypes. The mean values of the phenotypic data used in this study are presented in (Table S1. 1). For each of the trait, best linear unbiased predictors (BLUPs) using their replicated data at each site were

generated using linear mixed models in R (R Core 2015). The restricted maximum likelihood (REML) method was used to estimate variance components assuming a random effect model using lme4 package in R (R Core 2015). BLUP values were estimated for each trait, by treating genotype and site as a random effect.

### Genotyping

GBS was used to genotype all the 103 genotypes in the training population and was conducted at the Cornell University Institute of Genomic Diversity. Green leaf samples were collected early in the morning from the CFTs, freeze-dried for 3 days and stored in waterproof aluminum sachets. The freeze-dried samples were then shipped to ADNid laboratories in France for DNA extraction and quantification using the DNeasy 96 Plant Kit (QIAGEN). High-quality DNA was then sent to Cornell University's Institute of Genomic Diversity for genotyping using GBS. A multiplexed, high-throughput GBS procedure was conducted according to the procedure of Elshire et al. (2011). Sequence data were obtained from 96-plex Illumina HiSeq2000 runs. For genomic complexity reduction, the PstI restriction endonuclease was used.

A total of 155 billion base pair of good barcoded raw DNA sequence data were generated in GBS, with an average of 2 million reads per genotype. TASSEL UNEAK SNP calling algorithms (version 5.2.48) was used to determine SNP polymorphism, resulting in 82,254 SNPs. Nature Source Improved Plants (NSIP) applied an inhouse SNP calling algorithms to further filter to leave a high quality 2779 SNP dataset by decreasing error rate and increasing reliability (Professor Steve Tanksley, Pers. com, May 2016, NSIP). The SNP markers were then recoded as - 1, 0 and 1, corresponding to homozygous minor alleles, heterozygous and homozygous major alleles, respectively. Individuals with not more than 20% missing SNPs were selected and missing SNPs were imputed using EM algorithms in R using the A.mat function in the rrBLUP package (R Core 2015). A total of 2779 SNPs from the 103-tea genotypes were used in the present study.

## Relationship between the genotypes

All the statistical analysis was done in R (R Core 2015). To visualize the relatedness and population structure among the 103 genotypes, the realized genomic relationship matrix was created from the genotype matrix using the “A. mat” function in R via the rrBLUP (Endelman 2011). The kinship matrix for the pedigrees was estimated using the GeneticsPed package in R (Gorjanc et al. 2007). Principle component analysis (PCA) was determined using the 2779 SNP markers and was estimated using the k-means clustering function in R and the first two principle components were plotted (R Core 2015).

## Heritability estimation

Variance components and broad-sense heritability were estimated on an entry mean basis using the restricted maximum likelihood method (REML) with all factors set as random effects, using the ASReml-R version 4 package (Gilmour et al. 2015). Broad-sense heritability was calculated as the ratio of total genetic variance to total phenotypic variance. In multi-location trial analysis, broad-sense heritability was calculated as;

$$\frac{\sigma^2_g}{(\sigma^2_g + \sigma^2_{ge}/e + \sigma^2_{e/er})}$$

where  $\sigma^2_g$  is the genotypic variance component,  $\sigma^2_{ge}$  is the GxE variance component,  $\sigma^2_e$  is the residual variance and e and r are the number of environments and replicates within each environment, respectively (Zhang et al. 2017b).

Genomic heritability ( $h^2_g$ ) was estimated based on variance components estimated using the mixed model (de los Campos et al. 2015). Genetic variance was calculated as proportion of variance explained by regressing markers on phenotypes. The model was fitted in ASReml-R (Butler et al. 2009). Genomic heritability was estimated as;

$$\frac{\sigma^2_g}{(\sigma^2_g + \sigma^2_e)}$$

where  $\sigma^2_g$  is the genotypic variance component and  $\sigma^2_e$  is the residual variance.

## Prediction models

Six GS models namely, Bayesian Ridge regression best linear unbiased predictor (BRR) (Endelman 2011; Meuwissen et al. 2001), GBLUP (Endelman 2011), BayesA (Meuwissen et al. 2001), BayesB, (Meuwissen et al. 2001), BayesC (Meuwissen et al. 2001) and reproducing kernel Hilbert space (RKHS) regression (de los Campos and Pérez-Rodríguez 2016). The three RKHS models that we implemented were; (1) RKHS markers (RKHS-M) that involved using the G-matrix calculated from markers, (2) RKHS-pedigree (RKHS-P) that involved using the pedigree relationship matrix which was obtained from the pedigree and was twice the coefficient of ancestry, and (3) RKHS markers and pedigree (RKHS-MP) with the marker and pedigree relationship matrices as two kernels, where the additive effect was captured by regression on the markers and also with the (co)variance relationship derived from the pedigree.

## Prediction accuracies

The GBLUP model was performed using the “mixed.solve” function from the rrBLUP package (Endelman 2011). The other models were implemented using the BGLR package with default settings for priors (de los Campos and Pérez-Rodríguez 2016) in R version 4.0.3 (R Core 2015). The GS analysis in BGLR was set for 12,000 iterations and a burn-in setting of 2000. The predictive accuracy of all the GS models was estimated using a 5-fold cross-validation approach for all the traits. The data was randomly divided into 5 subsections, and one subset was also used as a distinct validation set (corresponding to 20% of the genotypes), while the remaining four groups (80% of all the genotypes) were used as training population for fitting the GS models. This process was repeated, each time with another subset, until all subsets had been used in both training and validation steps. Each analysis was repeated with 10 different cross-validation groupings and the mean GEBVs for the 10 subsets was calculated. The accuracy of the GS models was estimated as the Pearson correlation between the mean GEBVs and the observed phenotypes (biochemical traits);  $r(GEBV:y)$ .

## Training population size

This study used the genomic best linear unbiased prediction (GBLUP) model to test the effects of training population size (TPS) on genomic prediction accuracy. The GBLUP model was implemented using the mixed.solve function of the rrBLUP package in R (R Core 2015). Four levels of TPS (i.e., 20, 50, 80, and 90) were considered to evaluate the prediction accuracy of all the twelve traits. Similarly, the predictive accuracy was estimated using a 5-fold cross-validation approach, as the Pearson correlation between the biochemical BLUEs (best linear unbiased estimates) and its prediction from the GBLUP model.

## Results

### Descriptive statistics

All the quality traits for all the 103 genotypes were analyzed using NMR spectroscopy. The mean (mg per gram) biochemical contents, coefficient of variation and ranges are presented in Table 1. The coefficients of variation ranged from 10.1 to 56.5 %, signifying broad phenotypic variation. ANOVA revealed highly significant differences ( $p < 0.001$ ) among all the traits, signifying existence of genetic variation that can be exploited for breeding (Table 2).

## Relationship between the genotypes

The degree of relatedness of the genotypes based on the markers and pedigrees are shown in the heat map in Figs. 1 and 2. Values of the marker matrix are composed of both negative and positive values. The negative relationships are explained from the centering of the marker covariates, leading to centering of the entire marker-based matrix such that the sum of all elements in the matrix is zero. Negative values in the marker-based relationship matrix imply that the detection of an allele in one genotype makes it less likely to be detected in the other genotype, zero indicate absence of dependence, while positive values indicate an increased likelihood of an allele being detected in the other genotype.

There were two clear population structures as observed from the two heat maps (Figs. 1, 2). This was also confirmed by the principal component analysis (PCA) of the genotype data, with the first two principal components explaining 30% and 11%, respectively of the total marker variation, making a total of 41% (Fig. 3). The first two principal components were used because they explained the most variation.

**Table 1** Mean biochemical values (mg per gram), coefficient of variation (CV), and maximum and minimum values of the biochemical traits (mg per gram) across the sites

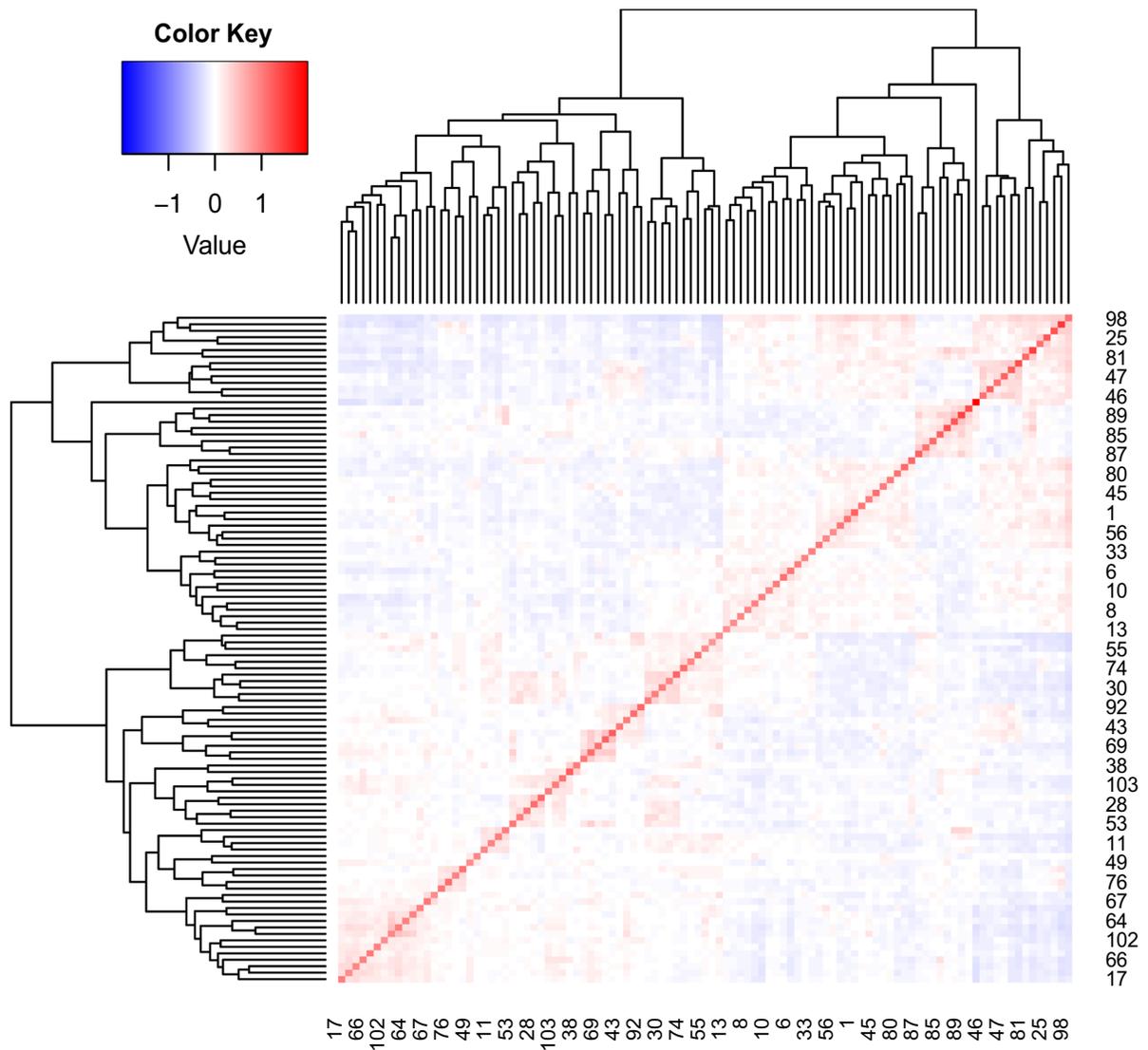
Trait	CV (%)	Mean (mg per gram)	Maximum (mg per gram)	Minimum (mg per gram)
Caffeine	10.1	37.83	46.66	27.92
Catechin	56.5	6.15	12.89	2.93
EC	19.6	28.81	34.95	22.51
ECG	13.7	60.65	112.61	40.51
EGC	18.9	43.42	69.30	17.13
EGCG	14.7	87.29	120.82	59.89
Gallic acid	22.8	4.08	5.82	2.94
GC	18.4	37.87	63.28	22.36
GCG	18.1	19.97	33.01	13.65
Theanine	29.9	19.09	29.82	10.70
Theobromine	38.7	4.28	6.84	2.10
Theogallin	18.5	15.83	24.39	11.47

**Table 2** Mean squares of ANOVA for biochemical compounds across the two sites

SOV	TB	Caffeine	Theogallin	GA	EC	GCG	EGC	EGC	EGCG	Catechin	GC	Theanine
Genotypes(G)	8.21***	77.82***	33.02***	2.1648***	44.59**	66.51***	977.99***	809.26***	844.8***	30.2***	616.04***	94.18***
Environment (E)	160.77***	433.98***	406.83***	28.65***	341.05***	18.3 ns	1623.19***	558.94**	3936.8***	0.42*	13.25 ns	666.04***
Interaction (GxE)	2.1 ns	13.93 ns	5.54 ns	0.69 ns	16.83 ns	14.44 ns	86.29 ns	63.27 ns	136.7 ns	10.5 ns	88.41***	38.09 ns
P values	5.309E-07	2.20E-16	2.20E-16	2.73E-10	0.016952	2.00E-16	2.20E-16	2.20E-16	2.20E-16	1.06E-10	2.00E-16	2.33E-12
Error	2.74	14.48	8.56	0.86	31.87	13.06	68.83	67.26	163.8	12.09	48.67	32.59
LSD ( $p < 0.05$ )	2.66	6.107	4.7	1.49	9.06	5.8	13.316	13.16	20.54	5.58	11.2	9.163

ns not significant, GCG gallocatechin gallate, ECG epicatechin gallate, EGC epigallocatechin gallate, EGCG Epigallocatechin gallate, GC, gallocatechin

\*Significant at  $p < 0.05$ ; \*\*significant at  $p < 0.01$ ; \*\*\*significant at  $p < 0.001$



**Fig. 1** Heat map of the marker-based relationship matrix of the 103 tea (*C. sinensis*) genotypes

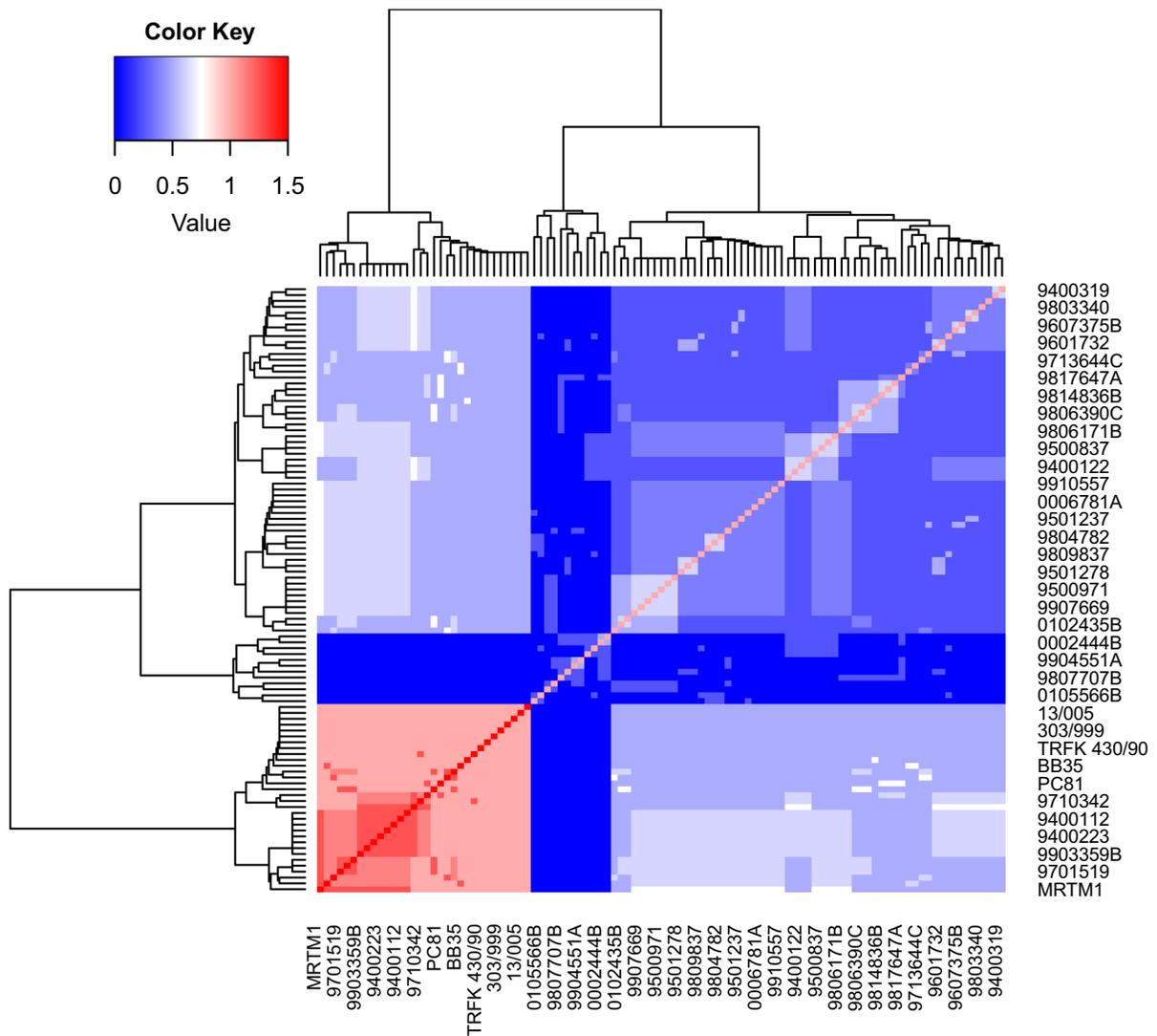
### Heritability

Broad sense heritability ranged from ECG (0.92) to EGCG (0.35). Traits with high broad sense heritability were ECG (0.92), EC (0.90), caffeine (0.82), EGC (0.81) and GC (0.81) (Fig. 4). Traits with low broad sense heritability were EGCG (0.35) and Theanine (0.56 (Fig. 4). Genomic heritability ranged from 0.99 (Theogallin) to 0.52 (EC) (Fig. 4). Traits with high genomic heritabilities were theogallin (0.99), ECG (0.99), theobromine (0.95), EGC (0.92) and EGCG (0.92) (Fig. 4). Traits with low genomic heritability were EC (0.52) and theanine (0.59) (Fig. 4). All traits

except EC, GC and GCG had a higher genomic heritability compared to broad sense heritability (Fig. 4).

### Prediction accuracies

RKHS-P had the lowest prediction for all the traits except GA. For theobromine, the models with the highest prediction accuracies were BRR (0.65) (Fig. 5). BayesB (0.51) had the highest prediction accuracy for caffeine, while RKHS-P (0.20) had the lowest prediction accuracy for the same trait (Fig. 5). RKHS-MP (0.68) had the highest prediction

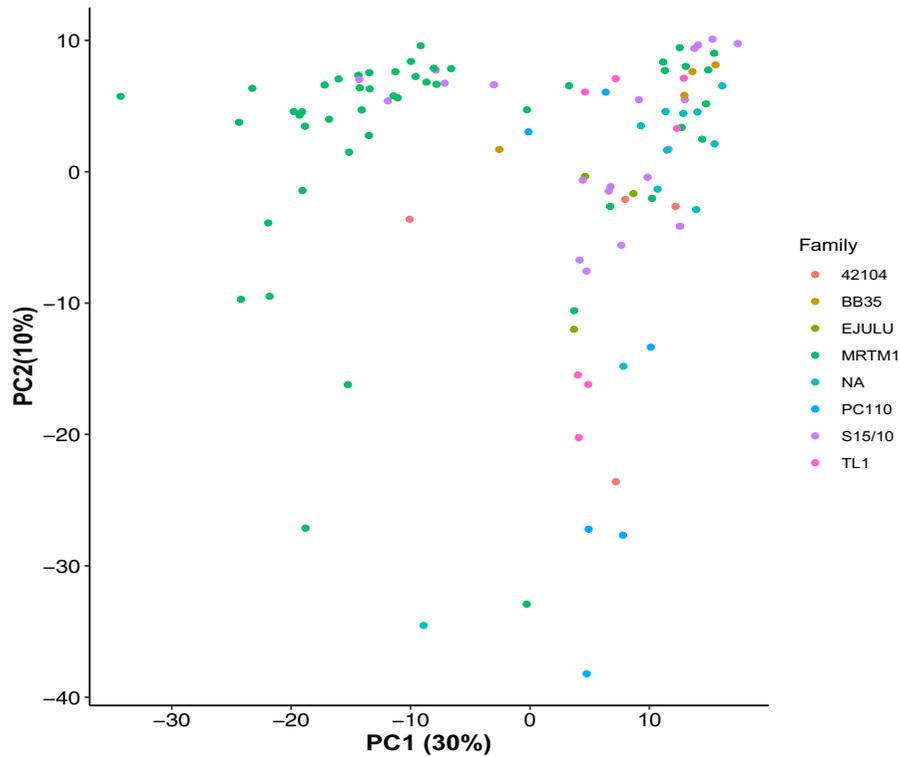


**Fig. 2** Heat map of the pedigree-based relationship matrices 103 tea genotypes illustrating the kinship between the individuals

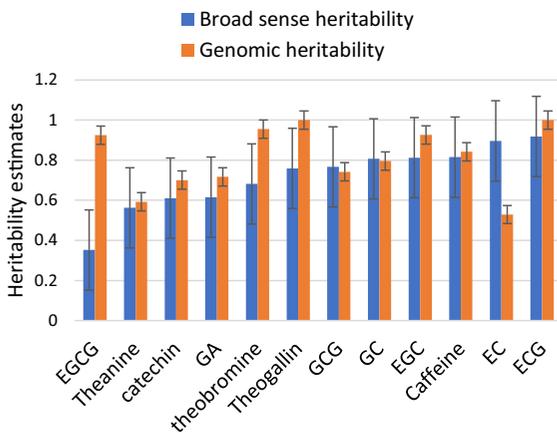
accuracies for theogallin, while RKHS-P (0.23) had the lowest. BRR (0.56) and BayesA (0.56) had the highest prediction accuracy for GA, while BayesB (0.33) had the lowest. GBLUP (0.54) and BayesA (0.53) had the highest prediction accuracies for EC, while BRR, BayesB and BayesC had similar prediction accuracies. For GCG, BRR (0.60) had the highest prediction accuracies. RKHS-MP (0.71), BRR(0.67) and BayesC (0.66) had the highest prediction accuracies for ECG. For EGC, GBLUP (0.64) recorded the highest prediction accuracy, while BRR, BayesA and BayesC had similar prediction accuracies for the same trait. For EGCG, BayesB (0.61), BayesA (0.59),

BayesC (0.59) and GBLUP (0.59) recorded the highest prediction accuracies. BRR (0.55) had the highest prediction accuracy for catechin. GBLUP, BayesA, BayesB, BayesC, RKHS-M and RKHS-MP had similar prediction accuracies for catechin. For GC, BayesB (0.49) had the highest prediction accuracy. The model with the highest prediction accuracy for theanine were RKHS-MP (0.44) (Fig. 5).

The mean prediction accuracies of the traits were averaged for all the GS models and the traits with the highest prediction accuracy were Theogallin (0.59), ECG (0.56) and Theobromine (0.54) (Fig. 5). Traits with the lowest mean prediction accuracies were



**Fig. 3** A PCA plot of the 2779 SNP markers for the 103 genotypes

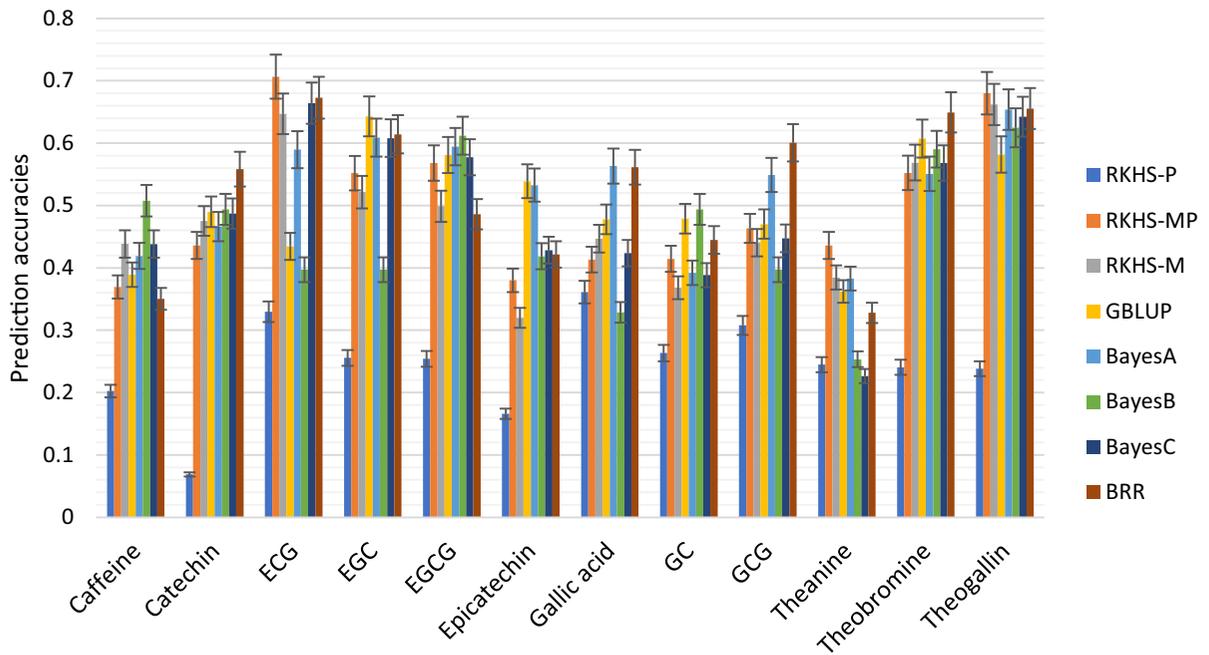


**Fig. 4** Comparison of broad sense and genomic heritability

Theanine (0.32) and caffeine (0.39). Similarly, the mean GS accuracies for all the traits was calculated. The performance of all the GS models were almost the same, with BBR (0.53), BayesA (0.52), GBLUP (0.50) and RKHS-MP (0.50) performing slightly better than the other models. BayesB had the lowest prediction accuracy in majority of the traits.

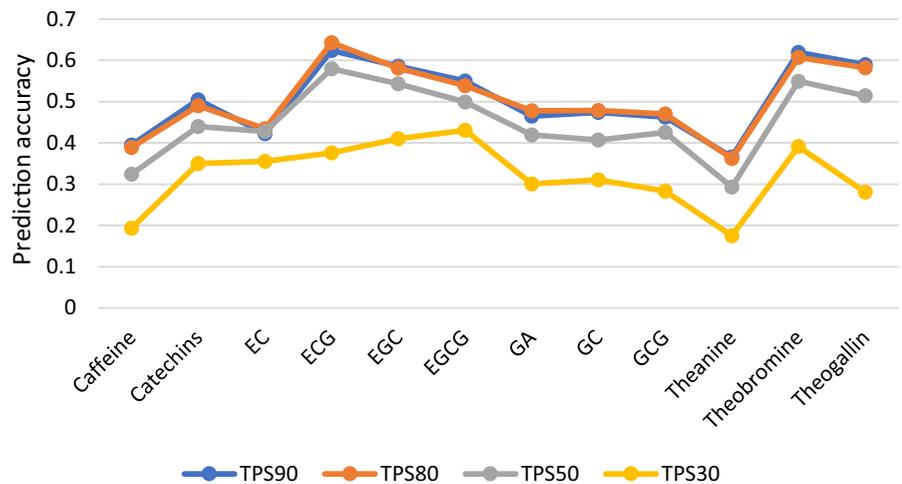
**Effect of training population size on prediction accuracy**

Prediction accuracy increased as the TPS increased for all the trait (Fig. 6). Comparing TPS30 with TP90, prediction accuracies increased from 0.37 to 0.64 for ECG (the most heritable trait), from 0.39 to 0.61 for theobromine, and from 0.41 to 0.59 for EGC. For EGCG, prediction accuracies increased from 0.43 to 0.54, while for EC, prediction accuracies increased from 0.36 to 0.43. For caffeine, prediction accuracies increased from 0.19 to 0.39, 0.35–0.49 for catechins, 0.28–0.58 for theogallin and 0.18–0.36 (Fig. 6). No significant differences between the mean accuracy of each training population size across traits were observed for TP90 and TP80, whereas accuracy for TP30 was significantly lower ( $p < 0.05$ ) compared to all other training population sizes (Fig. 6).



**Fig. 5** Mean accuracy of traits for the five studied GS models

**Fig. 6** Effect of training population size on accuracy of genomic selection for the quality traits



**Discussion**

Tea quality traits are difficult and expensive to measure, hindering the improvement of these traits using conventional breeding methods. GS is well suited for traits that are expensive and difficult to measure (Heffner et al. 2011), and therefore represents a promising approach for enabling cost-effective improvement of tea quality traits. In this study, we evaluated the potential of GS implementation to

increase genetic gain in tea breeding programmes. The impact of training population size and heritability on the prediction accuracy of twelve quality traits influencing tea quality were evaluated through cross-validations using GBLUP. The population used in this study consists of tea genotypes with diverse attributes. Known high-quality clones and poor-quality clones were included. The pedigree relationship matrices showed a higher relationship among the genotypes

than the marker-based matrices, because it does not account for Mendelian sampling.

#### Effect of heritability and training population size on genomic prediction accuracy

Generally, moderate to high prediction accuracies were observed for all the traits, and this could be attributed to the high heritability estimates observed. Similar findings have been reported in different crops by Ankamah-Yeboah et al. (2020), Mageto et al. (2020), Zhang et al. (2017b) and Arojju et al. (2020). The high prediction accuracies reported in our study shows that GS can be used in tea breeding to improve tea quality. The heritability of a trait significantly affects the response to selection and improves the efficiency of GS over phenotypic selection (Hayes et al. 2009; Zhang et al. 2017a). High heritability leads to increased gain from selection for the traits of interest (de los Campos et al. 2015; Kruijer et al. 2015).

Overall, genomic heritability estimates were higher than broad sense heritability for most traits, suggesting that higher genetic gains can be achieved by using molecular markers in tea breeding. The genomic heritability is the proportion of phenotypic variance explained by the regressing phenotypes on molecular markers. Many polymorphic markers are required to accurately estimate relatedness especially for distant relatives. GBLUP relies on estimating the realized kinship and is more accurate in estimating the hereditary relationships among genotypes (de Roos et al. 2009). Heritability of a trait could be improved by increasing the number of replications, years of recording phenotypic data and experimental sites (Zhang et al. 2017a).

Several GS methods have been developed for predicting complex traits and they include GBLUP, Bayesian alphabets (BayesA, BayesB and BayesCp), Ridge Regression (RR) BLUP, Random Forest and Support Vector Machine and deep learning (Cossa et al. 2017; Lorenz et al. 2011a). We compared six GS models characterized by two different assumptions with respect to the distribution of variance for marker effects. In RRBLUP, marker effects are equally shrunk and normally distributed with the same variance. Bayesian models allow marker-specific variances, and hence allow unequal shrinkage of marker effects. Koech et al. (2020) studied genome-enabled

prediction models for black tea quality and drought tolerance traits in discovery and validation populations. They only studied machine learning models, and although they showed promising results, a limited number of markers (1,421 DArTseq) were used. At the time of writing of this paper, there was no reported studies of genomic selection in tea using mixed parametric or semi-parametric models.

Our results showed that the GBLUP model performed similar to RKHS-M for all traits except ECG and EC. RKHS is a semi-parametric method where the genomic relationship matrix used in GBLUP is replaced by a kernel matrix, which enables nonlinear regression in a higher-dimensional feature space (Gianola et al. 2006). Several studies have reported that non-parametric models perform better than parametric models because they capture both additive and non-additive effects (e.g., dominance, epistasis). They can predict phenotypes better than the parametric models, especially where non-additive effects are important (Lebedev et al. 2020). For instance, in eucalyptus, RKHS had slightly better predictive abilities than four other models for traits with lower heritabilities (i.e. trunk CBH, height, and volume), but had the lowest prediction accuracies for pulp yield (Tan et al. 2017). In our study, RKHS-M did not differ in accuracy from the parametric methods. This agreed with other studies (Chen et al. 2018b; Juliana et al. 2017) who reported similar results. Cossa et al. (2013) compared GBLUP with the RKHS-M in maize and they concluded that there was no clear superiority of either of the models, although the RKHS-M performed slightly better than the GBLUP.

We also observed that RKHS-P model had the lowest prediction accuracies compared to the marker-based models for all traits. Similar results were also reported by Wolc et al. (2011) that marker-based methods had higher accuracies than the pedigree-based method. Likewise, Spindel et al. (2015) reported that marker based GS models were more superior to the pedigree-based prediction in rice for yield, height, and flowering time. The use of G-matrix has several benefits in genomic selection including (1) it can differentiate sibs and can also avoid selecting closely related sibs together, (2) it performs better when the pedigree information is not accurate or missing and (3) it can correct for pedigree errors (Juliana et al. 2017). However, the pedigree model had reasonable prediction accuracies for all traits, and this was because

Unilever Tea Kenya maintains accurate pedigree recording system and the families selected were small.

Although the RKHS-MP model performed well in most of the traits, and had the highest accuracies in ECG and theogallin, it did not perform significantly better than BRR and GBLUP. Several other studies (Crossa et al. 2013; Juliana et al. 2017), have reported higher prediction accuracies by using both pedigrees and markers in GS studies. While including markers and pedigree could improve the accuracy of selecting traits in tea breeding programmes, the benefits are not huge.

In forest trees, results for most traits showed similar prediction accuracies for RRBLUP and Bayesian models (Grattapaglia et al. 2018). For instance, Chen et al. (2018b) reported similar prediction accuracies in four genomic prediction models (GBLUP), Bayesian ridge regression (BRR), Bayesian LASSO(BL) and reproducing kernel hilbert space (RKHS) in Norway spruce (*Picea abies* (L.)). Similarly, Isik et al. (2016) observed similar predictive accuracies in maritime pine (*Pinus pinaster* Aiton) for GBLUP, BRR and BL prediction models. Tan et al. (2017) and Grattapaglia et al. (2018) proposed rrBLUP and GBLUP as the best models for use in forest tree breeding because they are computationally easy to use.

Increasing the training population size increased prediction accuracies across all measured traits but tended to plateau between TPS90 and TP80. Increasing number of genotypes at this point did not give any additional prediction accuracy. Increasing TPS increases accuracy by improving the estimation of marker effects (Heffner et al. 2011). Lozada et al. (2019) observed a positive correlation between TPS and prediction accuracy for yield and agronomic traits in soft red winter wheat. Similar results were also reported by Zhang et al. (2017b) in maize and Olatoye et al. (2020) in *Miscanthus* (grass). From our results for cross-validations, an optimal number of genotypes (~ 80 % of the entire population) should be included in the training panel to achieve improved predictions in tea. Beyond this, increasing TPS might not be longer advantageous for increasing accuracy.

### Implementing GS in tea breeding

The main factors that could be considered before implementing GS in a tea breeding programme include prediction models, the size of the training

population, the relationship between the training and the breeding populations, heritability, genetic architecture of the trait of interest in tea, marker density and cost-effective genotyping platforms.

The training population used to construct GS model should be closely related to the breeding population and should be large as possible as this improves the accuracy of estimating marker effects (Lorenz et al. 2011b). Zhang et al. (2017a) showed that prediction accuracy increased for all the traits in maize with increasing training population size. Since tea has a high allelic diversity, the training population should consist of genotypes with broad genetic diversity for the traits of interest.

Trait heritability is a key factor that significantly impacts on the accuracy of genomic selection (Heffner et al. 2011). Our findings agreed with previous studies that prediction accuracy increases with an increase in trait heritability (Zhang et al. 2017a). However, heritability could be improved by designing field experiments for the training population to increase the number of replications, testing sites and years of data collection (Mackay et al. 1999).

The density and type of markers to be used in constructing GS models influence the prediction accuracy (Goddard and Hayes 2011). In this study, SNP markers were used because they are abundant in the plant genome and they give higher prediction accuracies compared to other markers (Kwong et al. 2017). Cheaper options of SNP genotyping include GBS, a simple highly-multiplexed next generation sequencing platform that generates large numbers of SNPs (Elshire et al. 2011). GBS is less expensive compared to other platforms and can provide genome-wide marker coverage for species that lack a reference genome (Davey et al. 2011). However, SNP markers obtained by GBS usually contain a large proportion of missing data across samples because fragments of the genome are sequenced at low depth, and hence some loci could have zero coverage (Elshire et al. 2011). In GS, using a large number of markers and selecting a suitable imputation algorithm enables the use of low-density SNP markers without a major loss in prediction accuracy (Habier et al. 2009; Mulder et al. 2012). The most common imputation algorithms that could be used include; mean, singular value decomposition (SVD), traditional k nearest neighbor (kNN), expectation maximization (EM) and random forest regression imputation algorithms (Marchini and Howie

2010; Rutkoski et al. 2013). GS requires genome wide markers that explain most genetic variation (Meuwissen et al. 2001). Therefore an increase in the length of LD or in marker number steadily improves the prediction accuracy (Asoro et al. 2011).

The type of model used for GS could impact on the prediction accuracy and mainly depend on the complexity of the trait (Crossa et al. 2017). The main GS models developed differ on assumptions of the trait architecture and they include RRBLUP, GBLUP, Reproducing Kernel Hilbert Spaces(RKHS), Bayesian models (BayesA, BayesB, BayesC, BayesLASSO) and machine learning (Lorenz et al. 2011a; Wang et al. 2018). A suitable model could be tested and selected based on the complexity of the trait.

### Applying GS in tea improvement

Several limitations could affect the genetic gain of a GS programme in tea. A proper implementation of GS in tea breeding requires the optimization of field trial management and agricultural practices, and accurate phenotyping and genotyping of the training population. Generally, our results suggest that GS has a great potential in predicting superior tea quality genotypes. The main challenge facing all tea breeding programmes is the long generation interval, as it takes between 3 and 6 years for tea to grow from seedling to flowering (Mondal 2014). This means that developing an improved tea variety using conventional methods requires many years of field selection (Corley and

Tuwei 2018). GS in tea breeding could be beneficial by reducing the selection cycle time as shown in Fig. 7. This could be done by first applying GS early at the nursery stage. The genotypes with high GEBVs could be selected, tested in the field and the promising ones released for commercial planting. Compared to conventional field selection method, GS can improve genetic gain per unit time significantly.

### Conclusions

The evaluation of complex traits in tea such as quality using phenotypic selection is a difficult and expensive process using the standard conventional breeding process. Our results showed that the differences in prediction accuracies between the methods evaluated were small. Generally, BRR, BayesA, GBLUP and RKHS-MS models slightly outperformed the other methods. However, BRR and GBLUP could be preferred because they are computationally simple to use. Prediction accuracies increased with the increase in heritability and training population size. The high GS accuracies for nearly all the traits from our results clearly demonstrates the potential of GS using genome wide SNP markers to predict high quality varieties in a tea breeding programme. While the main benefit of GS in tea breeding is expected to be the reduction of the breeding cycle length by several years, the use of a realized genomic relationship matrix also enables the precise evaluation of genetic relationships and heritabilities. The next step would be to simulate a cost-benefit analysis to study the implications of manipulating the number of markers for cost-effective GS.

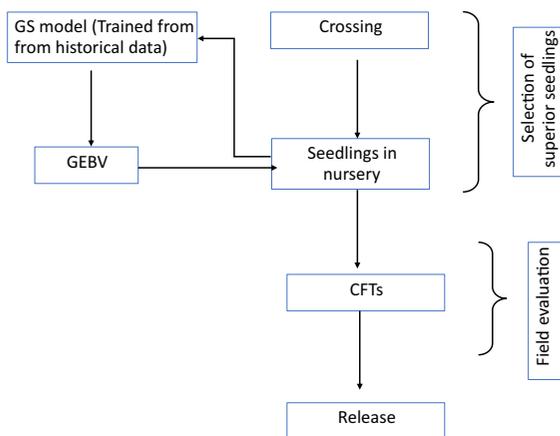
**Acknowledgements** This research was supported by Unilever R&D Colworth, University of Nottingham Malaysia and Unilever Tea Kenya.

**Author's contribution** N.L., F.M. and S.M. designed the research and wrote the manuscript. N.L. contributed to the phenotyping of the tea population, analyzed the genotypic data and constructed the G.S. models.

### Compliance with ethical standards

**Conflict of interest** All the authors declare that they have no conflict of interest.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any



**Fig. 7** Structure of a tea breeding scheme that aggressively uses genomic prediction to select improved seedlings for advanced field testing at the CFT stage

medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Ankamah-Yeboah T, Janss LL, Jensen JD, Hjortshøj RL, Rasmussen SK (2020) Genomic selection using pedigree and marker-by-environment interaction for barley seed quality traits from two commercial breeding programs. *Front Plant Sci.* <https://doi.org/10.3389/fpls.2020.00539>
- Arojju SK, Cao M, Jahufer MZZ, Barrett BA, Faville MJ (2020) Genomic predictive ability for foliar nutritive traits in perennial ryegrass. *G3 Genes Genomes Genet* 10:695–708
- Asoro FG, Newell MA, Beavis WD, Scott MP, Jannink J-L (2011) Accuracy and training population design for genomic selection on quantitative traits in Elite North American Oats. *Plant Genome J.* <https://doi.org/10.3835/plantgenome2011.02.0007>
- Bassi FM, Bentley AR, Charmet G, Ortiz R, Crossa J (2016) Breeding schemes for the implementation of genomic selection in wheat (*Triticum* spp.). *Plant Sci* 242:23–36. doi:<https://doi.org/10.1016/j.plantsci.2015.08.021>
- Beaulieu J, Doerksen T, Clément S, MacKay J, Bousquet J (2014) Accuracy of genomic selection models in a large population of open-pollinated families in white spruce *Heredity* 113:343
- Borse BB (2012) Novel bio-chemical profiling of Indian Black Teas with reference to quality parameters. *J Microb Biochem Technol.* <https://doi.org/10.4172/jbb.S14-004>
- Butler DG, Cullis BR, Gilmour AR, Gogel BJ, Thompson R (2009) ASReml-R reference manual (version 3). Department of Primary Industries, Brisbane
- Cerrudo D et al (2018) Genomic selection outperforms marker assisted selection for grain yield and physiological traits in a maize doubled haploid population across water treatments. *Front Plant Sci.* <https://doi.org/10.3389/fpls.2018.00366>
- Chen Q, Chen M, Liu Y, Wu J, Wang X, Ouyang Q, Chen X (2018a) Application of FT-NIR spectroscopy for simultaneous estimation of taste quality and taste-related compounds content of black tea. *J Food Sci Technol* 55:4363–4368. doi:<https://doi.org/10.1007/s13197-018-3353-1>
- Chen Z-Q et al (2018b) Accuracy of genomic selection for growth and wood quality traits in two control-pollinated progeny trials using exome capture as the genotyping platform in Norway spruce. *BMC Genom* 19:946. doi:<https://doi.org/10.1186/s12864-018-5256-y>
- Corley RHV, Tuwei G (2018) The well-Bred tea Bush. In: Carr M (ed) *Advances in tea agronomy*. Cambridge University Press, Cambridge, pp 106–136
- Crossa J et al (2013) Genomic prediction in maize breeding populations with genotyping-by
- Crossa J et al (2010) Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. *Genetics* 186:713–724
- Crossa J et al (2017) Genomic selection in plant breeding: methods, models, and perspectives. *Trends Plant Sci* 22:961–975
- Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat Rev Genet* 12:499
- de los Campos G, Pérez-Rodríguez P (2016) BGLR: Bayesian generalized linear regression R package version 1
- de los Campos G, Sorensen D, Gianola D (2015) Genomic heritability: What is it? *PLoS Genet* 11:e1005048
- de Roos APW, Hayes BJ, Goddard ME (2009) Reliability of genomic predictions across multiple populations. *Genetics* 183:1545–1553. <https://doi.org/10.1534/genetics.109.104935>
- El-Dien OG, Ratcliffe B, Klápště J, Chen C, Porth I, El-Kassaby YA (2015) Prediction accuracies for growth and wood attributes of interior spruce in space using genotyping-by-sequencing. *BMC Genomics* 16:370
- Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE* 6:e19379
- Endelman JB (2011) Ridge regression and other kernels for genomic selection with R package rrBLUP. *Plant Genome* 4:250–255
- Gianola D, Fernando RL, Stella A (2006) Genomic-assisted prediction of genetic value with semiparametric procedures. *Genetics* 173:1761–1776
- Gilmour AR, Gogel BJ, Cullis BR, Welham S, Thompson R (2015) ASReml user guide release 4.1 structural specification. VSN International Ltd, Hemel Hempstead
- Goddard ME, Hayes BJ (2011) Using the genomic relationship matrix to predict the accuracy of genomic selection. *J Anim Breed Genet* 128:409–421
- Gorjanc G, Henderson DA, Henderson MD, Runit S (2007) Package ‘GeneticsPed’
- Grattapaglia D et al (2018) Quantitative genetics and genomics converge to accelerate forest tree breeding. *Front Plant Sci* 9:66
- Habier D, Fernando RL, Dekkers JCM (2009) Genomic selection using low-density marker panels. *Genetics*
- Habier D, Fernando RL, Kizilkaya K, Garrick DJ (2011) Extension of the Bayesian alphabet for genomic selection. *BMC Bioinform* 12:186
- Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME (2009) Invited review: Genomic selection in dairy cattle: Progress and challenges. *J Dairy Sci* 92:433–443. doi:<https://doi.org/10.3168/jds.2008-1646>
- Heffner EL, Jannink J-L, Iwata H, Souza E, Sorrells ME (2011) Genomic selection accuracy for grain quality traits in biparental wheat populations. *Crop Sci* 51:2597–2606

- Isik F et al (2016) Genomic selection in maritime pine. *Plant Sci* 242:108–119
- Jannink JL, Lorenz AJ, Iwata H (2010) Genomic selection in plant breeding: from theory to practice briefings in functional. *Genomics* 9:166–177. <https://doi.org/10.1093/bfqp/elq001>
- Juliana P et al (2017) Genomic and pedigree-based prediction for leaf, stem, and stripe rust resistance in wheat. *Theor Appl Genet* 130:1415–1430
- Koech RK, Malebe PM, Nyarukowa C, Mose R, Kamunya SM, Loots T, Apostolides Z (2020) Genome-enabled prediction models for black tea (*Camellia sinensis*) quality and drought tolerance traits. *Plant Breed*
- Kruijer W et al (2015) Marker-based estimation of heritability in immortal populations. *Genetics* 199:379–398
- Kwong QB et al (2017) Genomic selection in commercial perennial crops: applicability and improvement in oil palm (*Elaeis guineensis* Jacq. *Sci Rep* 7:2872. <https://doi.org/10.1038/s41598-017-02602-6>
- Lawless HT, Heymann H (2010) Sensory evaluation of food: principles and practices. Springer, Berlin
- Le Gall G, Colquhoun IJ, Defernez M (2004) Metabolite profiling using 1H NMR spectroscopy for quality assessment of green tea, *Camellia sinensis* (L.). *J Agric Food Chem* 52:692–700
- Lebedev VG, Lebedeva TN, Chernodubov AI, Shestibratov KA (2020) Genomic selection for forest tree improvement: methods, achievements and perspectives. *Forests* 11:1190
- Lee J, Chambers DH (2007) A lexicon for flavor descriptive analysis of green tea. *J Sens Stud* 22:256–272
- Lorenz AJ et al (2011) Genomic selection in plant breeding. *Adv Agron* 6:77–123. <https://doi.org/10.1016/b978-0-12-385531-2.00002-5>
- Lorenz AJ et al (2011b) Genomic selection in plant breeding: knowledge and prospects. In: *Advances in agronomy*, vol 110. Elsevier, Amsterdam, pp 77–123
- Lozada DN, Mason RE, Sarinelli JM, Brown-Guedira G (2019) Accuracy of genomic selection for grain yield and agronomic traits in soft red winter wheat. *BMC Genet* 20:82. doi:<https://doi.org/10.1186/s12863-019-0785-1>
- Mackay IJ, Mackay IJ, Caligari PDS, Gibson JP (1999) Accelerated recurrent selection *Euphytica* 105:43–51. doi:<https://doi.org/10.1023/a:1003428430664>
- Mageto EK et al (2020) Genomic prediction with genotype by environment interaction analysis for kernel zinc concentration in tropical maize germplasm. *G3 Genes Genomes Genet* 10:2629. <https://doi.org/10.1534/g3.120.401172>
- Marchini J, Howie B (2010) Genotype imputation for genome-wide association studies. *Nat Rev Genet* 11:499
- Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157:1819–1829
- Mondal TK (2014) Breeding and biotechnology of tea and its wild species. Springer, Berlin
- Mulder HA, Calus MPL, Druet T, Schrooten C (2012) Imputation of genotypes with low-density chips and its effect on reliability of direct genomic values in Dutch Holstein cattle. *J Dairy Sci* 95:876–889
- Müller BSF et al (2019) Independent and Joint-GWAS for growth traits in Eucalyptus by assembling genome-wide data for 3373 individuals across four breeding populations. *New Phytol* 221:818–833
- Olatoye MO et al (2020) Training population optimization for genomic selection in *Miscanthus*. *G Genes Genomes Genet* 10:2465. <https://doi.org/10.1534/g3.120.401402>
- R Core Team (2015) R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2015. R Foundation for Statistical Computing
- Rutkoski JE, Poland J, Jannink J-L, Sorrells ME (2013) Imputation of unordered markers and the impact on genomic selection accuracy G3: Genes, Genomes. *Genetics* 3:427–439
- Spindel J et al (2015) Genomic selection and association mapping in rice (*Oryza sativa*): effect of trait genetic architecture, training population composition, marker number and statistical model on accuracy of rice genomic selection in elite, tropical rice breeding lines. *PLoS Genet* 11:e1004982
- Sverrisdóttir E et al (2017) Genomic prediction of starch content and chipping quality in tetraploid potato using genotyping-by-sequencing. *Theor Appl Genet* 130:2091–2108
- Tan B, Grattapaglia D, Martins GS, Ferreira KZ, Sundberg B, Ingvarsson PK (2017) Evaluating the accuracy of genomic prediction of growth and wood traits in two Eucalyptus species and their F 1 hybrids. *BMC Plant Biol* 17:110
- 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>
- Wolc A et al (2011) Breeding value prediction for production traits in layer chickens using pedigree or genomic relationships in a reduced animal model. *Genet Select Evol* 43:5
- Zhang A et al (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* 8:1916–1916. <https://doi.org/10.3389/fpls.2017.01916>
- Zhang A et al (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* 8:66. <https://doi.org/10.3389/fpls.2017.01916>
- Zheng X-Q, Li Q-S, Xiang L-P, Liang Y-R (2016) Recent advances in volatiles of teas. *Molecules* 21:338

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.