Genomic selection in maritime pine

Fikret Isik, Jérôme Bartholomé, Alfredo Farjat, Emilie E. Chancerel, Annie A. Raffin, Léopoldo Sanchez, Christophe Plomion, Laurent Bouffier

Genomic selection in maritime pine

Fikret Isik [1,2], Jérôme Bartholomé [1,3], Alfredo Farjat [2,4], Emilie Chancerel [1,3], Annie Raffin [1,3], Leopoldo Sanchez [5], Christophe Plomion [1,3], Laurent Bouffier [1,3*]

1/ INRA, UMR1202, BIOGECO, Cestas F-33610, France

2/ Permanent address: Department of Forestry and Environmental Resources, North Carolina State University, Raleigh, NC, USA

3/ Univ. Bordeaux, BIOGECO, UMR1202, Talence F-33170, France

4/ Department of Statistics, North Carolina State University, Raleigh, NC, USA

5/ INRA, UR0588, AGPF, 45075 Orléans, France

* Corresponding author:
Laurent Bouffier,
Address: INRA, UMR1202, BIOGECO, Cestas F-33610, France
Email: bouffier@pierroton.inra.fr

Abstract

A two-generation maritime pine (*Pinus pinaster* L.) breeding population (n=661) was genotyped using 2,500 SNP markers. The extent of linkage disequilibrium and utility of genomic selection for growth and stem straightness improvement were investigated. The overall intra-chromosomal linkage disequilibrium was $r^2 = 0.01$. Linkage disequilibrium corrected for genomic relationships derived from markers was smaller ($r_V^2 = 0.006$). Genomic BLUP, Bayesian ridge regression and Bayesian LASSO regression statistical models were used to obtain genomic estimated breeding values. Two validation methods (random sampling 50% of the population and 10% of the progeny generation as validation sets) were used with 100 replications. The average predictive ability across statistical models and validation methods was about 0.49 for stem sweep, and 0.47 and 0.43 for total height and tree diameter, respectively. The sensitivity analysis suggested that prior densities (variance explained by markers) had little or no discernible effect on posterior means (residual variance) in Bayesian prediction models. Sampling from the progeny generation for model validation increased the predictive ability of markers for tree diameter and stem sweep but not for total height. The results are promising despite low linkage disequilibrium and low marker coverage of the genome (~1.39 markers/cM).

Key words:
Linkage disequilibrium; tree breeding, genomic relationship, Bayesian regression, *Pinus pinaster*

1 Introduction

Genomic selection (GS) is considered a paradigm shift in animal and plant breeding [1] and has the potential to revolutionize breeding of forest trees. GS aims to trace all the quantitative trait loci (QTL) controlling phenotype to predict genetic merit of individuals [2, 3]. GS relies on a large number of DNA markers that cover the whole genome to exploit the linkage disequilibrium (LD) between markers and any QTL. Theoretically, if the marker coverage is dense enough, all the QTL controlling a trait will be in LD with at least one marker [4]. Therefore, the success of GS depends on the effective population size and on the extent of LD between DNA markers and loci affecting complex traits [5]. In contrast to marker-assisted selection, prior information on the association between phenotypes and markers, the location of QTL on the genome and their relative effect on the phenotype are not prerequisites for GS. Advances in high throughput genotyping technologies [6–8] has made available a large number of DNA markers to animal and crop breeders [9–12]. As a result, the concept of GS has been widely used for cattle breeding since 2008 [13–16] and has been extended to other animal and plant breeding programs world-wide [11, 17–20]. However, GS is its infancy with forest trees.

There has been extensive coverage of statistical methods used in GS and they were classified into two groups [21]. In the first group, the *i*-th phenotypic outcome (y_i) is regressed on markers via the regression function $g(\mathbf{x}_i, \boldsymbol{\beta}) = \mathbf{x}_i'\boldsymbol{\beta}$, where $\mathbf{x}_i$ is a vector of marker covariates and $\boldsymbol{\beta}$ is the vector of regression coefficients [22]. Bayesian shrinkage methods [2], ridge regression [23] or Bayesian LASSO regression [21] are statistical methods that fall into this category. Such models allow prediction of individual marker effects. The second approach uses genomic relationships derived from markers in a mixed model framework for prediction of genomic breeding values [24, 25]. This method is frequently called Genomic Best Linear Unbiased Prediction (GBLUP) and is an appealing method for ease of computation because there is no need to predict the marker effects. The number of solutions from the model is reduced to the number of individuals. Empirical and simulation studies suggest that the statistical methods usually differ only marginally in the predictive accuracy of genomic estimated breeding values [2, 26–28].

Forest trees, particularly conifers subjected to breeding, have long (15 years or more) cycles of breeding and testing [29]. Breeding trees is logistically difficult to implement because of

their reproductive biology (late flowering), their large physical sizes and, notably, their late maturation for the phenotypic evaluation of economically important traits. Using markers for selection has long been promoted to reduce the cost and time of progeny testing [30]. Several proof-of-concept studies of genomic prediction in forest trees have been published in recent years based on small (<8k) number of SNP markers [28, 31–35]. Despite advances in developing genomic resources for forest trees [36–38] and promising results from proof-of-concept studies, no application of GS in tree breeding programs has been reported [30, 39]. In addition, large physical genome sizes of conifers [40] may pose a challenge to achieving the necessary dense marker coverage of genomes. For example, the genome size of maritime pine (*Pinus pinaster* L.), is estimated to be 24.5 Gb [41]. The first whole-genome shotgun assembly of loblolly pine suggests a genome size of 20.1 Gb [42]. Since forest trees are still relatively undomesticated and characterized by large effective population sizes, the extent of LD is expected to be very low in these outcrossing species [42]. For example, in loblolly pine (*Pinus taeda* L.), the average short distance LD (physical scale) based on 19 candidate genes decayed to less than $r^2 = 0.2$ within about 1500 base pairs [43]. In maritime pine the pattern of long distance LD (genetic scale) was examined over 12 chromosomes using 194 unrelated individuals and 2600 SNP markers with an average map distance of 1.4 cM between markers [44]. Authors reported complete lack of long distance LD.

GS success, however, not only depends on the extent of LD at any given time but also on its dynamics over recombination cycles. Simulation studies suggested that response of GS will decline after each generation because LD weakens after recombination takes place [3, 45]. Therefore, a very large number of markers are likely needed to cover the whole genome in conifers in order to develop reliable and stable prediction models across generations. In this study, we used a maritime pine breeding population to estimate the extent of long distance LD and develop genomic prediction models. This is the first genomic prediction proof-of-concept study for this species. The study is based on a breeding population from two successive generations of the breeding scheme. The objectives were two-fold: i/ carry out LD analysis for each linkage group while correcting for genetic relatedness in the population, and ii/ compare three statistical models genomic Best Linear Unbiased Prediction (GBLUP), Bayesian ridge regression and Bayesian LASSO for their efficiency in genomic predictions of growth and stem sweep (a measure of tree stem straightness), two important traits of the maritime pine breeding program.

2 Material and Methods

2.1 Breeding population and pseudo phenotypes

The maritime pine breeding program in southwestern France started in the 1960s with the phenotypic selection of 635 individuals (G0 population) from unimproved pine plantations [46]. Selected trees were grafted in clonal archives for breeding. Progeny from G0 trees were first obtained by collecting cones on selected trees in the forest (wind pollination with unknown male pollen) then by crosses between grafted copies, using different mating schemes. Progeny (n ≈ 100) from crosses were tested in replicated field trials to select the next generation population (G1 population). The breeding population will have completed three generations of breeding, testing and selection in 2020 (selection of G3 population). Stem sweep (distance between the tree stem and a vertical pole at 1.5 m above ground) was measured between age 7 and 12 years. Total height and tree diameter at 1.3 m above ground were measured between age 6 and 15 years. A meta-analysis consisting of 39 progeny trials, with more than 300,000 data points, was carried out to estimate breeding values (EBV) for stem sweep at age 8 years, total height and tree diameter at age 12 years using the Treeplan genetic evaluation system [47]. For the present study, 184 unrelated founders (G0 trees) and 477 G1 trees were genotyped (**Figure S1**). Among the 477 G1 selections, 355 selections have both parents identified in the G0 population. The 122 remaining selections have only their mother identified in the G0 population as they were selected in open-pollinated progeny trials. In total, there were 191 maternal half-sib families in this G1 population. The number of individuals per half-sib family ranged from 1 to 13 with an average of 2.5 individuals per half-sib family. Inbreeding coefficients were equal to zero in the two-generation breeding population because it was comprised of unrelated founders and their offspring generation.

We used EBV as pseudo phenotypes in genomic predictions. All EBV were based on progeny test data and pedigree derived additive genetic relationships with high accuracies, ranging from 0.67 to 0.99. By definition, the accuracy of EBV is the correlation between the true breeding values and the EBV [48]. The accuracy r is estimated as

$r = \sqrt{1 - (\frac{S^2}{1+F\sigma_A^2})}$ where S is the standard error of the EBV, F is the coefficient of inbreeding and σ_A^2 is the additive genetic variance [49]. EBV for total height and tree

diameter were highly correlated, whereas EBV of these two traits had weak or no correlation with EBV of stem sweep (**Figure 1**). Although the range of accuracies was not large, using EBV as phenotypes in genomic prediction may introduce bias and heterogeneity [50]. We then compared EBV and the de-regressed breeding values (dEBV) as pseudo phenotypes to estimate the effect on reliability of genomic predictions. The dEBV for individual *i* was obtained as $\hat{u}_i^* = \hat{u}_i/r_i$, where $\hat{u}_i$ is the EBV and r_i is the accuracy of EBV [50]. The resulting de-regressed breeding values were then weighted according to $w_i = (1-h^2)/[c + (1-r_i)/r_i]h^2$, where h^2 is the heritability of the trait and *c* is the proportion of variance not accounted for by the markers (assumed to be 50%) [50, 51].

2.2 Genotyping and LD analysis

We used a 12K Infinium SNP array (Illumina Inc., San Diego, CA, USA) described by [52] to genotype 661 trees. One-year old pine needles (diploid tissue) were harvested to extract DNA. The Infinium assay was used to recover 2,600 informative markers from the G0 population. In this study the same markers were assayed in the G1 population. Missing genotypic data points (3,265 or 0.19%) were imputed from the marginal allele distribution for each marker. In other words, missing genotypes were sampled from scored genotypes (0, 1 and 2) assuming the population was in Hardy-Weinberg equilibrium. Markers with minor allele frequency (MAF) below 5% were discarded (100 out of 2600). In total, 2,500 markers were used for genomic prediction and model comparison. A high level of heterozygosity was found in the population with an average value of 0.39 (±0.02) over all individuals (**Figure S2**).

LPmerge software [53] was used to produce a composite genetic linkage map based on five published [54–56] and two unpublished (kindly provided by MT Cervera) linkage maps. Markers were assigned to a genetic map position to analyze the extent of LD along 12 maritime pine chromosomes (**Figure S3**). Intra-chromosomal LD (r^2) between pairs of loci was estimated using the R package *synbreed* [57]. We also calculated the LD for each chromosome corrected for the relatedness in the population using the *LDcorSV* package in R [58]. The realized genomic relationship matrix (see details below) derived from 2,500 SNP markers was used for the estimation of unbiased LD (r_V^2).

2.3 Statistical analyses for genomic predictions

Genomic BLUP (GBLUP), Bayesian ridge regression and Bayesian LASSO regression models were used in estimation of genomic breeding values. Details of these three statistical models are given below. We used the *BLR* 1.3 [51] and *synbreed* 0.10-4 [57] packages in R 3.1.2 environment [59] as the analytical framework for data organization, visualization, summary and statistical analyses.

2.3.1 Genomic BLUP

In GBLUP the inverse of the genomic relationship matrix **G**, derived from the markers is used to predict genomic estimated breeding values (GEBV). The realized genomic relationship matrix was computed as in [60]:

$$\mathbf{G} = \frac{(\mathbf{W}-\mathbf{P})(\mathbf{W}-\mathbf{P})'}{2\sum_{i=1}^{q} p_i(1-p_i)} \qquad \text{(Eq. 1)}$$

where **W** is the allele-sharing matrix with *n* rows (total number of genotyped individuals) and *q* columns (total number of markers). The elements of **W** were set to be -1, 0, and 1 denoting homozygote, heterozygote, and the other homozygote, respectively. **P** is the $n \times q$ matrix of allele frequencies with the *i*-th column given by $2(p_i - 0.5)$, where p_i is the observed allele frequency of all genotyped subjects. The denominator of Eq. 1 scales **G** to be analogous to the additive genetic relationships matrix (**A**) derived from pedigree [60]. The model for GBLUP using matrix notation is given by:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{e} \qquad \text{(Eq. 2)}$$

where **y** is the $n \times 1$ vector of phenotypes or "pseudo-phenotypes", **X** is the $n \times p$ design matrix for the fixed effects, **b** is the $p \times 1$ vector of fixed effects, **Z** is a $n \times n$ incidence matrix for the random effects, **u** is the $n \times 1$ vector of random additive effects for the individuals, and **e** is the $n \times 1$ vector of random errors. The vector of random additive effects **u** and residuals **e** are assumed to be independent of each other and to follow a multivariate normal distribution of the form: $\mathbf{u} \sim N(\mathbf{0}, \mathbf{G}\sigma_u^2)$ and $\mathbf{e} \sim N(\mathbf{0}, \mathbf{I}_n\sigma_e^2)$, where **0** denotes a *n*-dimensional vector of zeros and $\mathbf{I}_n$ is the *n*-dimensional identity matrix. The residual variance covariance matrix $\mathbf{I}_n\sigma_e^2$ in this case is a diagonal matrix with elements on the diagonal representing the individuals' accuracy (or reliability). The mixed model equations [61] were

solved to obtain GEBV:

$$\begin{bmatrix} \mathbf{X}' \mathbf{X} & \mathbf{X}' \mathbf{Z} \\ \mathbf{Z}' \mathbf{X} & \mathbf{Z}' \mathbf{Z} + \mathbf{G}^{-1} \alpha \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}' \mathbf{y} \\ \mathbf{Z}' \mathbf{y} \end{bmatrix} \quad \text{(Eq. 3)}$$

The scalar α is defined as $\alpha = \sigma_e^2 / \sigma_u^2$, which is equal to $\alpha = \frac{\sigma_e^2}{\sigma_a^2} 2 \sum_{i=1}^{q} p_i(1 - p_i)$ where σ_e^2 is the residual variance, σ_u^2 is the additive genetic variance explained by each locus, σ_a^2 is the total genetic variance and p_i is the frequency of *i*-th allele [61]. Parameter σ_u^2 is the additive genetic variance of the GBLUP approach. The so-called mixed model equations (Eq. 3) are not different from previously developed animal models, except for the fact that the genomic relationship matrix **G** was substituted for the numerator relationship matrix **A**. The **Z** matrix in this case is the *n*-dimensional identity matrix $\mathbf{I}_n$, where *n* is again the number of individuals with phenotypic and genotypic information. Note that because of the BLUP framework, the system of equations can automatically solve for individuals who are genotyped but lack of phenotypic information.

2.3.2 Bayesian Ridge Regression

Ridge regression is a shrinkage regression method that was originally intended [62] to deal with the problem of high correlation among predictors in linear regression models of the form $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$, where **y** is the $n \times 1$ response vector, **X** denotes the $n \times p$ design matrix, $\boldsymbol{\beta}$ the $p \times 1$ vector of regression coefficients, and **e** the $n \times 1$ error vector. Ridge regression estimates are obtained by minimizing the penalized residual sum of squares [63]:

$$\widehat{\beta}_{\mathrm{R}} = \underset{\beta}{\operatorname{argmin}} \left\{ |\mathbf{y} - \mathbf{X}\boldsymbol{\beta}|^2 + \lambda \sum_{\mathrm{i}=1}^{p} \beta_{\mathrm{i}}^{2} \right\} \quad \text{(Eq. 4)}$$

where λ is the regularization parameter and needs to be tuned. The Bayesian approach to ridge regression in the context of genomic selection consists of assuming a Gaussian likelihood with independent and identically distributed marker effects. Thus, phenotype **y** is modeled as a function of the individual markers as follows:

$$\mathbf{y} = \mathbf{1}_n \mu + \mathbf{W}\mathbf{m}_{\mathrm{R}} + \mathbf{e} \quad \text{(Eq. 5)}$$

where **y** is the $n \times 1$ response vector of phenotypes, $\mathbf{1}_n$ denotes a $n \times 1$ vector of ones, μ is a scalar representing the intercept, **W** is the $n \times q$ genotype matrix for the $q \times 1$ vector of

marker effects $\mathbf{m}_{\mathrm{R}}$, and $\mathbf{e}$ is the $n \times 1$ vector of random errors that follow a multivariate normal distribution $\mathbf{e} \sim \mathrm{N}(\mathbf{0}, \mathbf{I}_n \sigma_{\mathrm{e}}^2)$. To complete the Bayesian model formulation, prior distributions to all model unknowns must be assigned. The R package BLR assigns a flat prior to the vector of fixed effects, thus $p(\mu) \propto 1$ [51]. The vector $\mathbf{m}_{\mathrm{R}}$ is assigned a multivariate normal prior distribution with a common variance to all effects, that is $\mathbf{m}_{\mathrm{R}} \sim \mathrm{N}\left(\mathbf{0}, \mathbf{I}_q \sigma_{\mathrm{m_R}}^2\right)$. This prior induces estimates that are the Bayesian equivalent to those obtained from ridge regression. Parameter $\sigma_{\mathrm{m_R}}^2$ denotes the unknown genetic variance contributed by each individual marker and is assigned a scaled inverse χ^2 distribution with degrees of freedom $\mathrm{df}_{\mathrm{m_R}}$ and scale $S_{\mathrm{m_R}}$, so $\sigma_{\mathrm{m_R}}^2 \sim \chi^{-2}(\mathrm{df}_{\mathrm{m_R}}, S_{\mathrm{m_R}})$ under the parameterization such that $\mathrm{E}\left[\sigma_{\mathrm{m_R}}^2\right] = S_{\mathrm{m_R}} / \left(\mathrm{df}_{\mathrm{m_R}} - 2\right)$. Finally, the residual variance is assigned a scaled inverse χ^2 prior distribution with degrees of freedom df_{e}, and scale parameter S_{e}, that is $\sigma_{\mathrm{e}}^2 \sim \chi^{-2}(\mathrm{df}_{\mathrm{e}}, S_{\mathrm{e}})$.

2.3.3 Bayesian LASSO Regression

Both models described above assume that markers explain the same amount of variance (infinitesimal model) and markers effects are shrunk toward the mean at the same level. This is a convenient assumption but it is far from reality [51]. Shrinkage of markers could be specific and may provide higher accuracy of predictions [2]. The LASSO method combines shrinkage and variable selection principles [63]. A Bayesian version of LASSO regression was introduced to take the advantage of Gibbs sampling [64]. The model has the form of $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$, which estimates are defined by:

$$\widehat{\beta}_{\mathrm{L}} = \underset{\beta}{\operatorname{argmin}} \left\{ |\mathbf{y} - \mathbf{X}\boldsymbol{\beta}|^2 + \lambda \sum_{i=1}^{p} |\beta_i| \right\}, \qquad \text{(Eq. 6)}$$

where $\mathbf{y}$ is the $n \times 1$ response vector, $\mathbf{X}$ denotes the $n \times p$ design matrix, $\boldsymbol{\beta}$ the $p \times 1$ vector of regression coefficients, $\mathbf{e}$ the $n \times 1$ error vector. Parameter λ is referred to as the regularization parameter. As λ approaches to zero, the vector $\widehat{\beta}_{\mathrm{L}}$ becomes the ordinary least squares solution. Like with ridge regression, the LASSO regression coefficients are shrunk toward zero. However in the LASSO some of the coefficients are shrunk to exactly zero, which reduces the complexity of the model. As a result, and unlike ridge regression, the LASSO approach can also be used as a variable selection tool. The Bayesian LASSO regression model in the context of genomic selection is given by:

$$\mathbf{y} = \mathbf{1}_n\mu + \mathbf{W}\mathbf{m}_{\mathrm{L}} + \mathbf{e} \qquad \text{(Eq. 7)}$$

where $\mathbf{y}$ is the $n \times 1$ response vector of phenotypes, $\mathbf{1}_n$ denotes a $n \times 1$ vector of ones, μ is the intercept term, $\mathbf{W}$ is the $n \times q$ genotype matrix for the $q \times 1$ vector of marker effects $\mathbf{m}_{\mathrm{L}}$, and $\mathbf{e}$ is the $n \times 1$ vector of random errors that follow a multivariate normal distribution $\mathbf{e} \sim \mathrm{N}(\mathbf{0}, \mathbf{I}_n\sigma_{\mathrm{e}}^2)$, where σ_{e}^2 is the common error variance. In the BLR package, the vector of fixed effects is assigned a flat prior, hence $p(\mu) \propto 1$. The vector $\mathbf{m}_{\mathrm{L}}$ is given a multivariate normal distribution with marker-specific prior variances, that is, $\mathbf{m}_{\mathrm{L}} \sim \mathrm{N}(\mathbf{0}, \mathbf{T}\sigma_{\mathrm{mL}}^2)$, where $\mathbf{T} = \mathrm{diag}(\tau_1^2, \ldots, \tau_q^2)$. Parameters τ_j^2 are assigned independent and identically distributed exponential priors, namely $\tau_j^2 \sim Exp(\lambda^2)$ for $j = 1, \ldots, q$, where parameter λ^2 is given a Gamma prior distribution [64] with hyper-parameters r (shape) and δ (rate), thus $\lambda^2 \sim \mathrm{Gamma}(r, \delta)$. Finally, the residual variance is assigned a scaled inverse χ^2 prior distribution with degrees of freedom df_{e}, and scale parameter S_{e}. That is, the expectations of residuals are $\sigma_{\mathrm{e}}^2 \sim \chi^{-2}(\mathrm{df}_{\mathrm{e}}, S_{\mathrm{e}})$.

2.3.4 Model convergence and prior sensitivity analysis

For the Bayesian models, the convergence of each parameter was assessed from visual inspection of the corresponding trace plot of the Monte Carlo Markov Chains draws, the estimated density, and the autocorrelation function of the draws to make sure the chains were mixing well; that is successfully moving around the parameter space. Furthermore, we carried out the Gelman and Rubin diagnostic [65] with five chains generated from over-dispersed starting values, and also conducted the Geweke's diagnostic with the first 10% and last 50% of each chain to assess convergence. For the cross-validation analysis, 40,000 iterations with a 10,000 burn-in period were used.

The value of the hyper-parameters that define the Bayesian ridge and Bayesian LASSO models can be chosen to incorporate the analyst's prior belief or information about the proportion of phenotypic variance that is attributed to each component of the regression. Using the methods described in [51], we conducted a sensitivity analysis of the choice of prior on the resulting posterior distributions. The hyper-parameter values were set assuming that 20%, 40% and 60% of the observed phenotypic variance was due to genetic effects. For the validation analysis, the variance explained by the markers was assumed to be 50% for all traits.

2.3.5 Validation sampling methods and evaluation

Two validation methods were used to compare the predictive ability of the three different statistical models (GBLUP, Bayesian ridge and Bayesian LASSO). In the first method, the whole population was split randomly into two sets (regardless of generation): the training set (TS, *n=331*) and the validation set (VS, *n=330*). This sampling mimics the situation where only half of the population is phenotyped and genotyped, and the other half is genotyped but without phenotype to evaluate genomic selection efficiency. In the second method, G0 generation (184 individuals) and randomly sampled 90% of G1 population was used as training set (TS, 613). The remaining 10% of the G1 population was used as the validation set (VS, *n=48*). This scenario (G0+ 90% of G1) allowed having a larger training set (TS, *n=613*) and also allowed using the G0 population (founders) for model training. For both validation methods random sampling was repeated 100 times. The predictive ability of each replication was calculated with the validation set as the correlation between EBV and GEBV. The predictive ability and accuracy are related as $r(\hat{y}, y) = r(\hat{g}, g)H$, where H is the square root of broad-sense heritability [66]. The prediction bias of the fold was calculated by regressing the EBV on the GEBV in the validation set [26]. A regression coefficient of $b = 1$ suggests no bias, whereas $b < 1$ and $b > 1$ indicate inflation and deflation, respectively [57]. Spearman rank correlation of each fold within each replication was calculated between EBV and GEBV. Mean squared error of each fold was calculated between the EBV and GEBV of validation set as $\text{MSE}(EBV, GEBV) = n^{-1}\sum_{i=1}^{n}(EBV_i - GEBV_i)^2$. Individuals in validation set were ranked for their GEBV and the mean EBV of the top 10% individuals was calculated for each replicates.

3 Results

3.1 Extent of LD along the 12 maritime pine chromosomes

Among the 2,500 markers available for genomic prediction, 2,184 were mapped on the 12 linkage groups (LG) of the maritime pine genetic map (**Figure S3**). The number of markers on each chromosome ranged from 152 (LG#8) to 201 (LG#9). The overall intra-chromosomal LD was $r^2 = 0.011$. Relatedness corrected LD was smaller ($r_V^2 = 0.006$). Both LD estimates were significantly smaller than 0.1 (Pr < 0.01). Average regular LD (r^2) values for each LG ranged from 0.008 to 0.019 and from 0.005 to 0.012 for related adjusted LD (r_V^2) (**Table 1**). For both r^2 and r_V^2, LD values decayed rapidly with increasing genetic distance as illustrated for LG1 (**Figure 2**). Illustrations of LD for all 12 chromosomes are presented in a supplemental file (**Figure S4**). As expected, the correction for relatedness among genotypes using realized genomic relationships tended to decrease LD values especially for genetically closely linked markers. The scatter plot of regular LD and relationship corrected LD suggested that when genetic relationships are taken into account, the frequency of LD estimates greater than 0.6 was considerably reduced (**Figure S5**). Among the 876 r_V^2 values greater than 0.6, about 90.8% involved physically linked SNPs belonging to the same contig, while 5.6% involved completely genetically linked SNPs (0 cM in the component maps). The remaining 3.6% are more likely related to bias in composite linkage map construction as discussed in [44].

3.2 Genetic relationships

Additive genetic relationships derived from the pedigree and realized genomic relationships derived from SNP markers are presented in **Figure 3**. Concerning the pedigree-based estimates, a high majority of individuals had zero relationships as a result of limited relatedness in the population. Non-zero covariances clustered into two distinct groups, 0.25 and 0.50, corresponding to expectations for half and full-sibs, respectively. Relationships derived from shared alleles are also clustered into two groups when non-zero coefficients were excluded, but they showed a continuous bimodal distribution centered on 0.25 and 0.50 (**Figure 3)**. Inbreeding coefficients (*F*) were distributed around 0 within the expected range

of variation (**Table 2, Figure S6**). Since the base population is known to be unrelated, null inbreeding coefficients were expected for the progeny population (G1).

3.3 Model convergence and marker-trait association

The Gelman and Rubin convergence plots calculated from five chains were used to monitor model convergence for the Bayesian statistical models Bayesian ridge, and Bayesian LASSO (**Figure S7A, B**). The Gelman and Rubin convergence plots suggest that convergence of the Monte Carlo Markov Chains to the stationary distribution is reached after about 10,000 iterations since approximate convergence is diagnosed when the upper limit of the shrink factor is close to one.

The sensitivity analysis of the choice of prior on the resulting posterior distributions suggested that prior densities had little or no discernible effect on posterior means (**Figure S8A**). The hyper-parameter values that define the prior distributions were set assuming that 20%, 40% and 60% of the observed phenotypic variance were due to genetic effects. The posterior means (0.27, 0.26, 0.25) and associated standard deviations (≈ 0.03) were very similar to each other for the Bayesian ridge model. Similarly, for the Bayesian LASSO model, the posterior means (0.17, 0.16, 0.16) and corresponding standard deviations (≈ 0.02) were almost identical (**Figure S8B**).

3.4 De-regressed breeding values (dEBV) as pseudo-phenotypes

We compared model performance statistics for using dEBV and EBV as pseudo phenotypes. The comparisons were carried out for the cross-validation scenario of sampling 10% of the G1 population. It can be concluded that using the dEBV versus EBV did not have a considerable effect on the model performance statistics (**Table 3**, **Table S2**). For height and stem sweep the predictive ability estimates were similar for EBV and dEBV regardless of statistical models used. For example, the predictive ability for tree diameter was about 0.43 (average across three statistical models) when dEBV was used as a phenotype (**Table S2**), while it was 0.47 with EBV. For tree height and diameter, the rank correlations were slightly higher for dEBV (**Table S2**) than for EBV (**Table 3**) while they were identical (0.55) for stem sweep. Except for supplemental **Table S2**, all the results in this paper are based on using EBV as phenotype.

3.5 Genomic prediction and model comparison

The performance of statistical models were compared using the predictive ability, rank correlations, bias, mean-squared error of genetic predictions and the phenotype mean of the top 10% ranked selections based on GEBV (**Tables 3 and 4**). Statistical models produced similar predictive ability estimates (see an example for tree height in **Figure 4**). The predictive ability estimates of markers for diameter were low compared to height and stem sweep. For example, using GBLUP, the average predictive ability for diameter was 0.39 while it was 0.47 and 0.49 for height and stem sweep, respectively (**Table 4**). Predictive abilities of genomic predictions for height and stem sweep were similar regardless of sampling method and statistical model used (**Figure S9**). The average predictive ability for height across statistical models and validation scenarios was about 0.47, while it was 0.49 for stem sweep.

Rank correlations between GEBV and EBV were equal to or smaller than the predictive ability estimates for the three traits, but they followed the same trend, i.e. when a predictive ability value was high for a given validation scenario, so was the rank correlation. Statistical models had almost no effect on rank correlations. Diameter, with a range of 0.36 to 0.42, had lower rank correlations than tree height (range of means 0.43 to 0.46). Stem sweep exhibited higher rank correlations across different sampling and statistical models with a range of 0.47 to 0.55.

The bias of the model is the deviation of the slope from the regression line $b = 1$ when phenotype is regressed on GEBV. A model with no bias ($b = 1$) is preferred in genetic evaluation so that GEBV would be on the same scale as BV obtained from the phenotypic data. In terms of bias, splitting the population in half for model training and validation performed generally better (**Table 4**) than sampling 10% of G1 population for validation (**Table 3**). Bayesian LASSO always produced larger bias (slope < 1) estimates in both sampling methods than GBLUP and Bayesian ridge. Considering the best random sampling/model combination, bias were smaller than 0.05 (slope ≥ 0.95) for the three traits.

Interestingly, when the individuals were ranked based on their GEBV and the phenotypic mean of the top 10% were compared for statistical models, Bayesian LASSO provided clearly higher GEBV for the validation sets (**Table 3, 4**) and for the whole data sets (**Table**

S1). For example, when 50% random sampling was used for cross validation, the mean diameter of the 10% best individuals was 1.18 for Bayesian LASSO whereas it was 0.97 and 0.95 for Bayesian ridge and GBLUP models, respectively (**Table 4**). A similar large difference between statistical models was also observed when sampling 10% of the G1 population as validation set (**Table 3**).

4 Discussion

4.1 Cross validation over generations of the breeding population

In this study, we introduced the first comprehensive analysis of relatedness corrected LD and genomic prediction in maritime pine, an important species for wood production in the South-Western Europe. Our study covers a two-generation breeding population. This allowed us to sample the progeny generation (G1) for cross validation. Previous studies on forest trees split a population of the same generation into training and validation sets to predict GEBV [28, 32, 34]. Splitting the same generation for model training and validation may not be optimal to estimate the prediction ability of markers for forward selection. In such cases the sibs are expected to share large segments of chromosomes and a small number of markers might be constructing the pedigree by tracing large haplotypes segregated in the families [31] as opposed to exploiting the LD associated with factors controlling the traits. This is especially true for a population with a small effective population size and for families with large numbers of progeny. The predictive ability of models based on splitting the same generation for cross-validation will likely decrease substantially in the subsequent generations. Our results suggest that there is a marginal advantage for the across generation validation over that performed from random halves. This advantage could come from higher levels of relatedness between training and validation in the former approach, although the benefit of a larger training set cannot be excluded. In general, it has been suggested [67] that training populations should comprise large levels of diversity for producing a robust calibration, while the validation is best when it comprises relatedness links to training sets. In this sense, including our founder population in the training set helped to have a larger diversity, while validating over progeny assured relatedness links.

4.2 Extent of LD and consequences in terms of GS

LD is widely used to assess the effect of evolutionary forces (e.g. selection, genetic drift) on a population and to infer marker-trait association in genome-wide association studies [68]. The extent of LD and its relationship with genomic prediction accuracies have been widely covered in simulation studies [4, 69, 70]. In a simulated barley population, marker density and LD interactions caused significant changes in prediction accuracies and the estimation of

LD was suggested to determine marker density in genomic prediction [71]. When marker density is at low coverage, predictions may rely primarily on genetic relationships established by the markers rather than tracing LD between marker-tagged QTL [70, 71]. DNA markers can capture additive genetic relationships even with low LD and may produce genomic predictions with accuracies greater than zero [70]. By definition LD is a non-random association between markers. In a breeding population, LD might be biased because of the genetic relationships or because of the genetic structure (different genetic groups) in the population [58]. We carried out LD analysis for each maritime pine chromosome, based on a composite genetic linkage map [44]. The average LD across 12 LGs was about $r^2 = 0.006$ after correcting for genetic relatedness with a mean density of 1.39 marker/cM. Our results are consistent with previous studies. High LD estimates in maritime pine were reported for only physically linked SNP markers located on the same gene [44]. They reported a lack of long distance LD over 12 chromosomes and no inter-chromosomal LD. Absence of long range LD was also reported in *Pinus sylvestris* L. [72] and in *Pinus taeda* L [73]. Low LD suggests that a large number of markers might be needed to trace QTLs that are in LD with markers. Using deterministic simulations, 10 markers/cM genotyping density was suggested to achieve the same accuracy level obtained from a classical genetic evaluation based on phenotype [69]. In our study, the average marker coverage was about 1.39 marker/cM. Considering the large genome (24Gb) of maritime pine and the low LD, a larger number of markers would be needed to trace QTLs controlling complex traits for genomic prediction. However, the genetic size of conifer genomes is still on the range of any plant with much smaller physical sizes. Therefore, there are strong reasons to believe that some parts of conifer genomes are completely inactive with respect to recombination. This idea was reinforced by a recent study in *Cryptomeria japonica* (a member of the *Cupressaceae* botanical family) in which high LD was found along BAC clones, i.e. across large physical distances [74]. Besides, the recombination comparison of angiosperms and gymnosperms showed that gymnosperms have less recombinations than angiosperms [75]. This consideration should certainly be taken into account before drawing definite conclusions about marker density for genomic prediction in conifers.

4.3 The distribution of genomic relationships

Genomic relationships were used in estimation of unbiased LD and in GBLUP. Diagonal elements of the genetic relationship matrix (1+*F*) estimated from markers were distributed around 1.0 within the expected range of variance (**Table 2, Figure S6**). A fraction of realized genomic relationships derived from markers was negative suggesting that pairs of individuals shared fewer alleles than expected based on allele frequencies. Similarly, positive realized genomic coefficients close to zero indicate that pairs of individuals shared more alleles than expected based on marginal allele frequencies. This is due to the fact that genetic relationships estimated from markers are a function of observed allele frequencies. Genotyping mistakes and/or errors in pedigree may cause estimates outside of expected boundaries of relationships [76]. Nevertheless, realized genomic relationships had a continuous distribution with modes matching the expectations for half and full-sibs (**Figure 3**). This distribution of realized genomic relationships around expectations is the result of the Mendelian sampling term that creates differences in allele sharing between pairs of sibs for a given family [77]. A pedigree-based approach of genetic evaluation is blind to these differences, as all sibs are assumed to share the same genome that is identical by descent.

4.4 Trait heritability versus predictive ability of markers

Heritability of a trait is considered an important factor affecting the predictive ability of markers in genomic predictions [13, 69]. It is assumed that genomic predictive ability is particularly beneficial for traits with low heritability [78]. In loblolly pine, predictive ability of markers for lignin and cellulose was considerably higher than the predictive ability of markers for tree height and stem volume [31]. Cellulose and lignin content are known to have higher heritability than growth traits in loblolly pine [79]. In another study on loblolly pine a high positive correlation was observed between the heritability of 17 traits and predictive ability of markers [28]. In this study we analyzed tree diameter ($h^2 = 0.17$), stem sweep ($h^2 = 0.25$) and tree height ($h^2 = 0.30$) with different heritability estimates. The overall predictive ability of markers for stem sweep and tree height was higher compared to diameter, supporting the conclusions of an earlier simulation study [78] and also concurring with the results reported for loblolly pine [31,32]. For traits with low heritability, larger number of markers and greater model training size might be needed to obtain reliable and unbiased GEBV.

4.5 Overall prediction ability of markers

The average reliability of GEBV in this study varied from 0.43 to 0.55. These estimates are lower than the correlation of true and EBV (ranged from 0.67 to 0.99) obtained from progeny tests. This is not surprising since the training models had small sample size; marker density was low and the long range LD decayed sharply within a short genetic distance along the chromosomes. Yet the predictive abilities of markers in this study were similar or higher for growth traits (diameter = 0.43, height = 0.47) compared to previous studies on several conifers. For example, accuracy values of 0.38 to 0.49 were reported for growth traits in a cloned population of loblolly pine [28]. In white spruce (*Picea glauca* Voss.) low accuracies were reported for within family (0.36) and between family (0.18) selection for tree height at age 22 years [34].

4.6 Comparison of prediction accuracy between statistical models

In our study, the statistical models did not differ noticeably for various model evaluation statistics (predictive ability, rank correlations, mean squared error etc.). Similar results were reported in the literature based on simulations [2, 21]. In loblolly pine, the accuracy of predictions for 17 traits differed marginally across various statistical models [28]. In a simulation study that mimicked a barley population, a GBLUP approach produced more accurate predictions than other methods that fit markers simultaneously, suggesting that capturing the genetic relationship was more important than capturing LD. GBLUP is an attractive statistical method to obtain GEBV and has been widely used in simulation and empirical data [21]. In GBLUP, the experimental design factors can be included in the models. The genotype by environment interaction can be formulated and variance-covariance structures can be incorporated into GBLUP models to account for heterogeneity. Furthermore, multivariate models can be fit to account for genetic correlations between traits. Computationally GBLUP is less demanding because it does not predict the individual marker effects in a mixed model approach but this is also the major drawback of GBLUP.

4.7 EBV versus de-regressed EBV

In prediction of GEBV, the phenotype of individuals is regressed on genetic markers in a training population. The ideal phenotype would be true breeding values (TBV) measured in a

population of unrelated individuals without selection [50]. In reality the TBV are never known. Instead phenotypes adjusted for systematic effects (e.g. experimental design factors), estimated breeding values (EBV) from genetic evaluation, average progeny performance or repeated measures on individuals are used to estimate marker effects [50]. The consequence of using different phenotypes in animal breeding was a concern and various statistical methods were introduced to *de-regress* EBV [80]. If family is the target for selection and EBV of families are used in GEBV, the benefit of de-regressing the EBV may not materialize in forest trees as we saw in this study. The analyzed population consisted of mostly unrelated individuals from two generations of breeding. An important part of the individuals (306 out of 661) had at least one parent unknown (wind pollinated). The EBV of 661 individuals were derived from a large number of progeny (n≈100) with high reliability (ranging from 0.66 to 0.99). These considerations argue in favor of using EBV instead of de-regressed EBV.

4.8 Conclusions

In conclusion, this study showed encouraging results of applying genomic selection in maritime pine. The predictive ability of markers for growth and stem sweep was around 0.50 despite low average LD (0.006) observed in the population with one cycle of breeding. Sampling of progeny for model validation increased the accuracy GEBV. A larger number of individuals from third generation need to be genotyped and the whole population needs to be analyzed to further test the validity of the GS model. Such analysis is underway by our group.

Acknowledgement

This study was carried out with financial support from the European Union's Seventh Framework Programmes: Procogen (n° 289841) and Noveltree (n° 211868). EC and FI were supported by Noveltree and ProCoGen projects, respectively. We are also thankful to the North Carolina State University Cooperative Tree Improvement Program for partially funding the sabbatical stay of FI at INRA, Bordeaux, France. JB was supported by a Postdoctoral Fellowship from "Conseil Général des Landes". The authors are thankful to Unité expérimentale Forêt Pierroton, UE 0570, INRA, 69 route d'Arcachon, 33612 CESTAS and to GIS Groupe Pin Maritime du Futur for installing and measuring progeny tests. The Groupe Pin Maritime du Futur is supported by the Conseil Régional d'Aquitaine and by the Direction Régionale de l'Alimentation, de l'Agriculture et de la Forêt d'Aquitaine.

Authors' contributions:

FI analyzed data, wrote the first draft. JB analyzed data, helped in writing and editing. AF, helped in writing and editing, run Bayesian prior sensitivity analysis. LS, CP and LB helped in writing and interpretation of the results. EC and AR carried out the genotyping and phenotyping work, respectively. LB and CP conceived, designed and coordinated the study. All authors read and approved the final manuscript.

Literature cited

1. Pryce JE, Daetwyler HD: Designing dairy cattle breeding schemes under genomic selection: a review of international research. Anim Prod Sci 2012, 52:107–114.

2. Meuwissen THE, Hayes BJ, Goddard ME: Prediction of total genetic value using genome-wide dense marker maps. Genetics 2001, 157:1819–1829.

3. Goddard ME, Hayes BJ, Meuwissen THE: Genomic selection in livestock populations. Genet Res 2011, 92:413.

4. Hayes B, Goddard M: Genome-wide association and genomic selection in animal breeding. Genome 2010, 53:876–883.

5. Brito FV, Neto JB, Sargolzaei M, Cobuci JA, Schenkel FS: Accuracy of genomic selection in simulated populations mimicking the extent of linkage disequilibrium in beef cattle. BMC Genet 2011, 12:80.

6. Perkel J: SNP genotyping: six technologies that keyed a revolution. Nat Methods 2008, 5:447–453.

7. Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE: A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. PLoS ONE 2011, 6:e19379.

8. Kumar S, Banks TW, Cloutier S: SNP Discovery through Next-Generation Sequencing and Its Applications. Int J Plant Genomics 2012, 2012:e831460.

9. Matukumalli LK, Lawley CT, Schnabel RD, Taylor JF, Allan MF, Heaton MP, O'Connell J, Moore SS, Smith TPL, Sonstegard TS, Van Tassell CP: Development and Characterization of a High Density SNP Genotyping Assay for Cattle. PLoS ONE 2009, 4:e5350.

10. Ganal MW, Polley A, Graner E-M, Plieske J, Wieseke R, Luerssen H, Durstewitz G: Large SNP arrays for genotyping in crop plants. J Biosci 2012, 37:821–828.

11. Kranis A, Gheyas AA, Boschiero C, Turner F, Yu L, Smith S, Talbot R, Pirani A, Brew F, Kaiser P, Hocking PM, Fife M, Salmon N, Fulton J, Strom TM, Haberer G, Weigend S, Preisinger R, Gholami M, Qanbari S, Simianer H, Watson KA, Woolliams JA, Burt DW: Development of a high density 600K SNP genotyping array for chicken. BMC Genomics 2013, 14:59.

12. Yu H, Xie W, Li J, Zhou F, Zhang Q: A whole-genome SNP array (RICE6K) for genomic breeding in rice. Plant Biotechnol J 2014, 12:28–37.

13. Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME: Invited review: Genomic selection in dairy cattle: Progress and challenges. J Dairy Sci 2009, 92:433–443.

14. Eggen A: The development and application of genomic selection as a new breeding paradigm. Anim Front 2012, 2:10–15.

15. Boichard D, Guillaume F, Baur A, Croiseau P, Rossignol MN, Boscher MY, Druet T, Genestout L, Colleau JJ, Journaux L, Ducrocq V, Fritz S: Genomic selection in French dairy cattle. Anim Prod Sci 2012, 52:115.

16. Hayes BJ, Cogan NOI, Pembleton LW, Goddard ME, Wang J, Spangenberg GC, Forster JW: Prospects for genomic selection in forage plant species. Plant Breed 2013, 132:133–143.

17. Jannink JL, Lorenz AJ, Iwata H: Genomic selection in plant breeding: from theory to practice. Brief Funct Genomics 2010, 9:166–177.

18. Heffner EL, Lorenz AJ, Jannink J-L, Sorrells ME: Plant Breeding with Genomic Selection: Gain per Unit Time and Cost. Crop Sci 2010, 50:1681.

19. Poland JA, Brown PJ, Sorrells ME, Jannink J-L: Development of High-Density Genetic Maps for Barley and Wheat Using a Novel Two-Enzyme Genotyping-by-Sequencing Approach. PLoS ONE 2012, 7:e32253.

20. Wellmann R, Preu\s s S, Tholen E, Heinkel J, Wimmers K, Bennewitz J: Genomic selection using low density marker panels with application to a sire line in pigs. Genet Sel Evol 2013, 45:28.

21. de los Campos G, Hickey JM, Pong-Wong R, Daetwyler HD, Calus MPL: Whole-Genome Regression and Prediction Methods Applied to Plant and Animal Breeding. Genetics 2013, 193:327–345.

22. de los Campos G, Naya H, Gianola D, Crossa J, Legarra A, Manfredi E, Weigel K, Cotes JM: Predicting Quantitative Traits With Regression Models for Dense Molecular Markers and Pedigree. Genetics 2009, 182:375–385.

23. Endelman JB: Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. Plant Genome J 2011, 4:250.

24. Wiggans GR, Sonstegard TS, VanRaden PM, Matukumalli LK, Schnabel RD, Taylor JF, Schenkel FS, Van Tassell CP: Selection of single-nucleotide polymorphisms and quality of genotypes used in genomic evaluation of dairy cattle in the United States and Canada. J Dairy Sci 2009, 92:3431–3436.

25. VanRaden PM, Van Tassell CP, Wiggans GR, Sonstegard TS, Schnabel RD, Taylor JF, Schenkel FS: Invited Review: Reliability of genomic predictions for North American Holstein bulls. J Dairy Sci 2009, 92:16–24.

26. Luan T, Woolliams JA, Lien S, Kent M, Svendsen M, Meuwissen TH: The accuracy of genomic selection in Norwegian red cattle assessed by cross-validation. Genetics 2009, 183:1119–1126.

27. Gao H, Christensen OF, Madsen P, Nielsen US, Zhang Y, Lund MS, Su G: Comparison on genomic predictions using three GBLUP methods and two single-step blending methods in the Nordic Holstein population. Genet Sel Evol 2012, 44.

28. Resende MDV, Resende Jr MFR, Sansaloni CP, Petroli CD, Missiaggia AA, Aguiar AM, Abad JM, Takahashi EK, Rosado AM, Faria DA: Genomic selection for growth and wood quality in Eucalyptus: capturing the missing heritability and accelerating breeding for complex traits in forest trees. New Phytol 2012.

29. Mullin T, B Andersson, J-C Bastien, J Beaulieu, RD Burdon, WS Dvorak, JN King, T Kondo, J Krakowski, SJ Lee, SE McKeand, L Pâques, A Raffin, JH Russell, T Skr?ppa, M Stoehr, A Yanchuk: Economic Importance, Breeding Objectives and Achievements. In Genetics, Genomics and Breeding of Conifers. Science Publishers; 2011:40–127.

30. Isik F: Genomic selection in forest tree breeding: the concept and an outlook to the future. New For 2014, 45:379–401.

31. Zapata-Valenzuela J, Isik F, Maltecca C, Wegrzyn J, Neale D, McKeand S, Whetten R: SNP markers trace familial linkages in a cloned population of Pinus taeda—prospects for genomic selection. Tree Genet Genomes 2012, 8:1307–1318.

32. Zapata-Valenzuela J, Whetten RW, Neale DB, McKeand SE, Isik F: Genomic Estimated Breeding Values Using Genomic Relationship Matrices in a Cloned Population of Loblolly Pine. G3 GenesGenomesGenetics 2013.

33. Resende Jr MFR, Munoz P, Acosta JJ, Peter GF, Davis JM, Grattapaglia D, Resende MDV, Kirst M: Accelerating the domestication of trees using genomic selection: accuracy of prediction models across ages and environments. New Phytol 2011.

34. Beaulieu J, Doerksen T, Clément S, MacKay J, Bousquet J: Accuracy of genomic selection models in a large population of open-pollinated families in white spruce. Heredity 2014, 113:343–352.

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

36. Chancerel E, Lepoittevin C, Provost GL, Lin Y-C, Jaramillo-Correa JP, Eckert AJ, Wegrzyn JL, Zelenika D, Boland A, Frigerio J-M, Chaumeil P, Garnier-Géré P, Boury C, Grivet D, González-Martínez SC, Rouzé P, Peer YV de, Neale DB, Cervera MT, Kremer A, Plomion C: Development and implementation of a highly-multiplexed SNP array for genetic mapping in maritime pine and comparative mapping with loblolly pine. BMC Genomics 2011, 12:368.

37. Geraldes A, DiFazio SP, Slavov GT, Ranjan P, Muchero W, Hannemann J, Gunter LE, Wymore AM, Grassa CJ, Farzaneh N, Porth I, McKown AD, Skyba O, Li E, Fujita M, Klápště J, Martin J, Schackwitz W, Pennacchio C, Rokhsar D, Friedmann MC, Wasteneys GO, Guy RD, El-Kassaby YA, Mansfield SD, Cronk QCB, Ehlting J, Douglas CJ, Tuskan GA: A 34K SNP genotyping array for Populus trichocarpa: Design, application to the study of natural populations and transferability to other Populus species. Mol Ecol Resour 2013, 13:306–323.

38. Wegrzyn JL, Liechty JD, Stevens KA, Wu L-S, Loopstra CA, Vasquez-Gross HA, Dougherty WM, Lin BY, Zieve JJ, Martínez-García PJ, Holt C, Yandell M, Zimin AV, Yorke JA, Crepeau MW, Puiu D, Salzberg SL, Jong PJ de, Mockaitis K, Main D, Langley CH, Neale DB: Unique Features of the Loblolly Pine (Pinus taeda L.) Megagenome Revealed Through Sequence Annotation. Genetics 2014, 196:891–909.

39. Muranty H, Jorge V, Bastien C, Lepoittevin C, Bouffier L, Sanchez L: Potential for marker-assisted selection for forest tree breeding: lessons from 20 years of MAS in crops. Tree Genet Genomes 2014:1–20.

40. Mackay J, Dean JFD, Plomion C, Peterson DG, Cánovas FM, Pavy N, Ingvarsson PK, Savolainen O, Guevara MÁ, Fluch S, Vinceti B, Abarca D, Díaz-Sala C, Cervera M-T: Towards decoding the conifer giga-genome. Plant Mol Biol 2012, 80:555–569.

41. Chagné D, Brown G, Lalanne C, Madur D, Pot D, Neale D, Plomion C: Comparative genome and QTL mapping between maritime and loblolly pines. Mol Breed 2003, 12:185–195.

42. Neale DB, Wegrzyn JL, Stevens KA, Zimin AV, Puiu D, Crepeau MW, Cardeno C, Koriabine M, Holtz-Morris AE, Liechty JD, Martínez-García PJ, Vasquez-Gross HA, Lin BY, Zieve JJ, Dougherty WM, Fuentes-Soriano S, Wu L-S, Gilbert D, Marçais G, Roberts M, Holt C, Yandell M, Davis JM, Smith KE, Dean JF, Lorenz WW, Whetten RW, Sederoff R, Wheeler N, McGuire PE, et al.: Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. Genome Biol 2014, 15:R59.

43. Neale DB, Savolainen O: Association genetics of complex traits in conifers. Trends Plant Sci 2004, 9:325–330.

44. Plomion C, Chancerel E, Endelman J, Lamy J-B, Mandrou E, Lesur I, Ehrenmann F, Isik F, Bink MC, Bouffier L, others: Genome-wide distribution of genetic diversity and linkage disequilibrium in a mass-selected population of maritime pine. BMC Genomics 2014, 15:171.

45. Muir WM: Comparison of genomic and traditional BLUP-estimated breeding value accuracy and selection response under alternative trait and genomic parameters. J Anim Breed Genet Z Für Tierz Zücht 2007, 124:342–355.

46. Illy G: Recherches sur l'amélioration génétique du pin maritime. 1966, 23:757–948.

47. McRae T, Dutkowski G, Pilbeam D, Powell M, Tier B: Genetic evaluation using the TREEPLAN system. Charleston, SC, USA; 2004.

48. Mrode RA: Linear Models for the Prediction of Animal Breeding Values: 3rd Edition. CABI; 2014.

49. Gilmour AR, Gogel B, Cullis B, Thompson, R.: ASReml User's Guide. Release 3.0. Hemel Hempstead, UK: VSN International, Ltd.; 2009.

50. Garrick DJ, Taylor JF, Fernando RL: Deregressing estimated breeding values and weighting information for genomic regression analyses. Genet Sel Evol 2009, 41:55.

51. Pérez P, de los Campos G, Crossa J, Gianola D: Genomic-Enabled Prediction Based on Molecular Markers and Pedigree Using the Bayesian Linear Regression Package in R. Plant Genome J 2010, 3:106.

52. Chancerel E, Lamy J-B, Lesur I, Noirot C, Klopp C, Ehrenmann F, Boury C, Le Provost G, Label P, Lalanne C: High-density linkage mapping in a pine tree reveals a genomic region associated with inbreeding depression and provides clues to the extent and distribution of meiotic recombination. BMC Biol 2013, 11:50.

53. Endelman JB, Plomion C: LPmerge: an R package for merging genetic maps by linear programming. Bioinformatics 2014:btu091.

54. Chancerel E, Lamy J-B, Lesur I, Noirot C, Klopp C, Ehrenmann F, Boury C, Provost GL, Label P, Lalanne C, Léger V, Salin F, Gion J-M, Plomion C: High-density linkage mapping in a pine tree reveals a genomic region associated with inbreeding depression and provides clues to the extent and distribution of meiotic recombination. BMC Biol 2013, 11:50.

55. de Miguel M, de Maria N, Guevara MÁ, Diaz L, Sáez-Laguna E, Sánchez-Gómez D, Chancerel E, Aranda I, Collada C, Plomion C, others: Annotated genetic linkage maps of Pinus pinaster Ait. from a Central Spain population using microsatellite and gene based markers. BMC Genomics 2012, 13:527.

56. de Miguel M, Cabezas J-A, María N de, Sánchez-Gómez D, Guevara M-Á, Vélez M-D, Sáez-Laguna E, Díaz L-M, Mancha J-A, Barbero M-C, Collada C, Díaz-Sala C, Aranda I, Cervera M-T: Genetic control of functional traits related to photosynthesis and water use efficiency in Pinus pinaster Ait. drought response: integration of genome annotation, allele association and QTL detection for candidate gene identification. BMC Genomics 2014, 15:464.

57. Wimmer V, Albrecht T, Auinger H-J, Schön C-C: synbreed: a framework for the analysis of genomic prediction data using R. Bioinformatics 2012, 28:2086–2087.

58. Mangin B, Siberchicot A, Nicolas S, Doligez A, This P, Cierco-Ayrolles C: Novel measures of linkage disequilibrium that correct the bias due to population structure and relatedness. Heredity 2012, 108:285–291.

59. R Core Team: R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2013.

60. VanRaden PM: Efficient methods to compute genomic predictions. J Dairy Sci 2008, 91:4414–4423.

61. Henderson CR: Best Linear Unbiased Estimation and Prediction under a Selection Model. Biometrics 1975, 31:423.

62. Hoerl AE, Kennard RW: Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics 1970, 12:55–67.

63. Tibshirani R: Regression Shrinkage and Selection via the Lasso. J R Stat Soc Ser B Methodol 1996, 58:267–288.

64. Park T, Casella G: The Bayesian Lasso. J Am Stat Assoc 2008, 103:681–686.

65. Gelman A, Rubin DB: Inference from Iterative Simulation Using Multiple Sequences. Stat Sci 1992, 7:457–472.

66. Legarra A, Robert-Granié C, Manfredi E, Elsen J-M: Performance of Genomic Selection in Mice. Genetics 2008, 180:611–618.

67. Rincent R, Laloë D, Nicolas S, Altmann T, Brunel D, Revilla P, Rodríguez VM, Moreno-Gonzalez J, Melchinger A, Bauer E, Schoen C-C, Meyer N, Giauffret C, Bauland C, Jamin P, Laborde J, Monod H, Flament P, Charcosset A, Moreau L: Maximizing the Reliability of Genomic Selection by Optimizing the Calibration Set of Reference Individuals: Comparison of Methods in Two Diverse Groups of Maize Inbreds (Zea mays L.). Genetics 2012, 192:715–728.

68. Barton NH: Estimating multilocus linkage disequilibria. Heredity 2000, 84:373–389.

69. Grattapaglia D, Resende MDV: Genomic selection in forest tree breeding. Tree Genet Genomes 2011, 7:241–255.

70. Habier D, Fernando RL, Dekkers JCM: The Impact of Genetic Relationship Information on Genome-Assisted Breeding Values. Genetics 2007, 177:2389–2397.

71. Zhong S, Dekkers JCM, Fernando RL, Jannink J-L: Factors Affecting Accuracy From Genomic Selection in Populations Derived From Multiple Inbred Lines: A Barley Case Study. Genetics 2009, 182:355–364.

72. Kujala ST, Savolainen O: Sequence variation patterns along a latitudinal cline in Scots pine (Pinus sylvestris): signs of clinal adaptation? Tree Genet Genomes 2012, 8:1451–1467.

73. Eckert AJ, van Heerwaarden J, Wegrzyn JL, Nelson CD, Ross-Ibarra J, González-Martínez SC, Neale DB: Patterns of population structure and environmental associations to aridity across the range of loblolly pine (Pinus taeda L., Pinaceae). Genetics 2010, 185:969–982.

74. Moritsuka E, Hisataka Y, Tamura M, Uchiyama K, Watanabe A, Tsumura Y, Tachida H: Extended Linkage Disequilibrium in Noncoding Regions in a Conifer, Cryptomeria japonica. Genetics 2012, 190:1145–1148.

75. Jaramillo-Correa JP, Verdú M, González-Martínez SC: The contribution of recombination to heterozygosity differs among plant evolutionary lineages and life-forms. BMC Evol Biol 2010, 10:22.

76. Simeone R, Misztal I, Aguilar I, Legarra A: Evaluation of the utility of diagonal elements of the genomic relationship matrix as a diagnostic tool to detect mislabelled genotyped animals in a broiler chicken population. J Anim Breed Genet 2011, 128:386–393.

77. Wang H, Misztal I, Legarra A: Differences between genomic-based and pedigree-based relationships in a chicken population, as a function of quality control and pedigree links among individuals. J Anim Breed Genet 2014, 131:445–451.

78. Goddard M: Genomic selection: prediction of accuracy and maximisation of long term response. Genetica 2009, 136:245–257.

79. Sykes R, Isik F, Li B, Kadla J, Chang H-M: Genetic variation of juvenile wood properties in a loblolly pine progeny test. Tappi J 2003, 2:3–8.

80. Habier D, Fernando RL, Kizilkaya K, Garrick DJ: Extension of the Bayesian alphabet for genomic selection. BMC Bioinformatics 2011, 12:186.

Figure 1: Scatter plots (lower diagonal), histograms (diagonal) and correlations (upper diagonal) between stem sweep, tree height and diameter with probability values (H_0: $r = 0$).

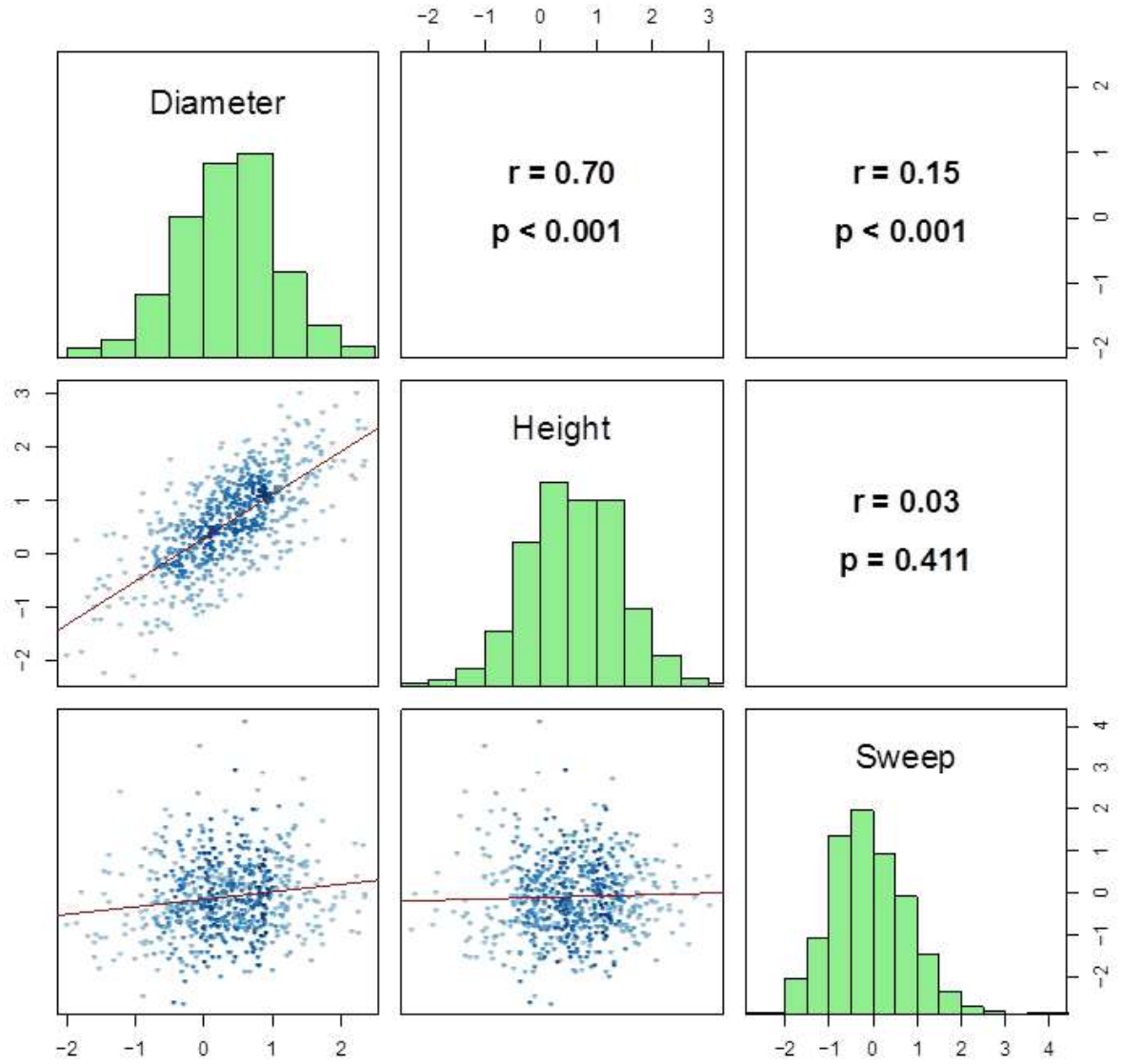
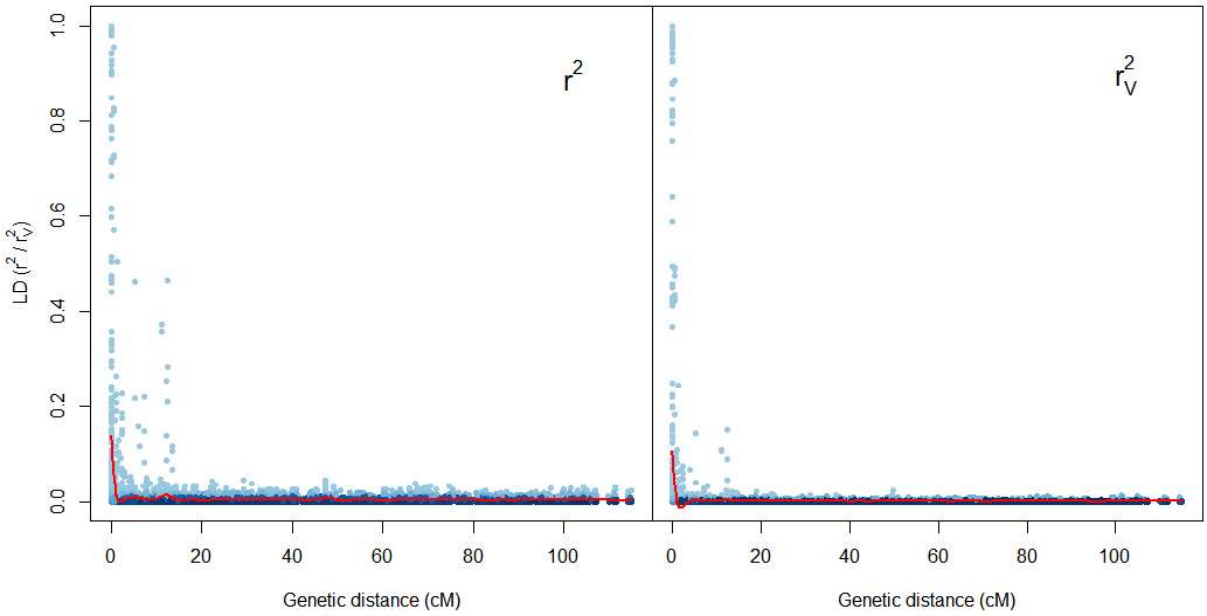


Figure 2: Linkage disequilibrium (LD) in chromosome 1 based on squared allele frequency (r^2) and LD corrected for genetic relatedness (r_V^2).

A) Regular (r^2) and corrected (r_V^2) LD between pairs of markers against the genetic distance (cM) as scatter plot with smoothed spline (red), showing a rapid decline of LD within a short genetic distance.



B) Regular (left) and corrected (right) LD between pairs of markers as heat map (the upper diagonal matrix of LD between markers). The lines on the diagonal show the location of the markers on linkage group #1.

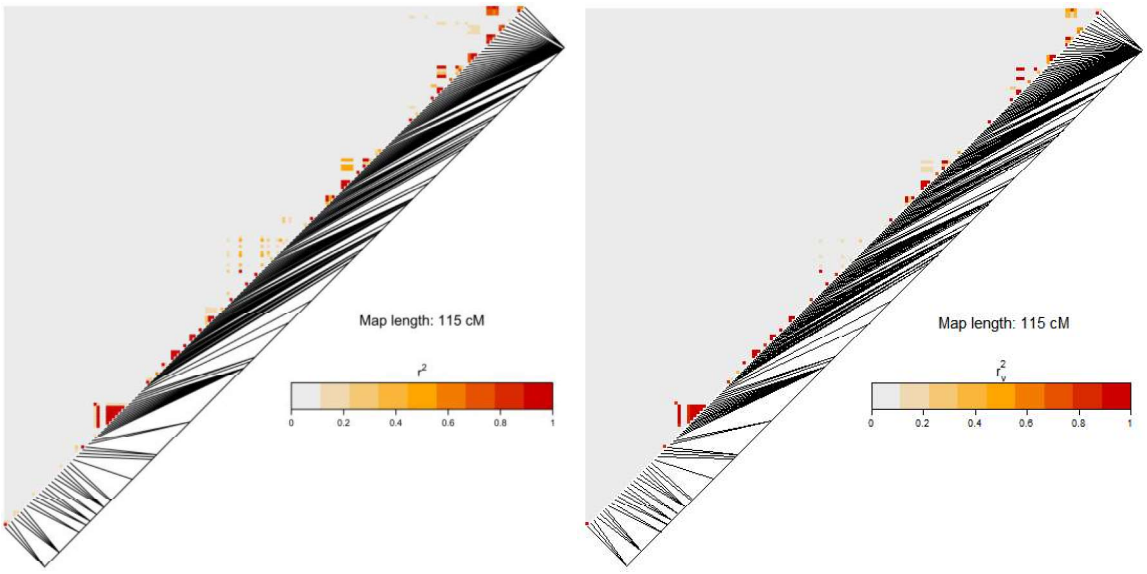


Figure 3: Additive genetic relationships derived from the pedigree (top panel) and realized genomic relationships (bottom panel) derived from SNP markers. A high majority of individuals had zero covariance (relationship) showing limited relatedness in the population. The distribution of the coefficients of relatedness is tri-modal, clearly showing three peaks (0, 0.25 and 0.50), the expected coefficients between half-sibs and full-sibs (or parent offspring). The scale of the y-axis is square root of the frequency.

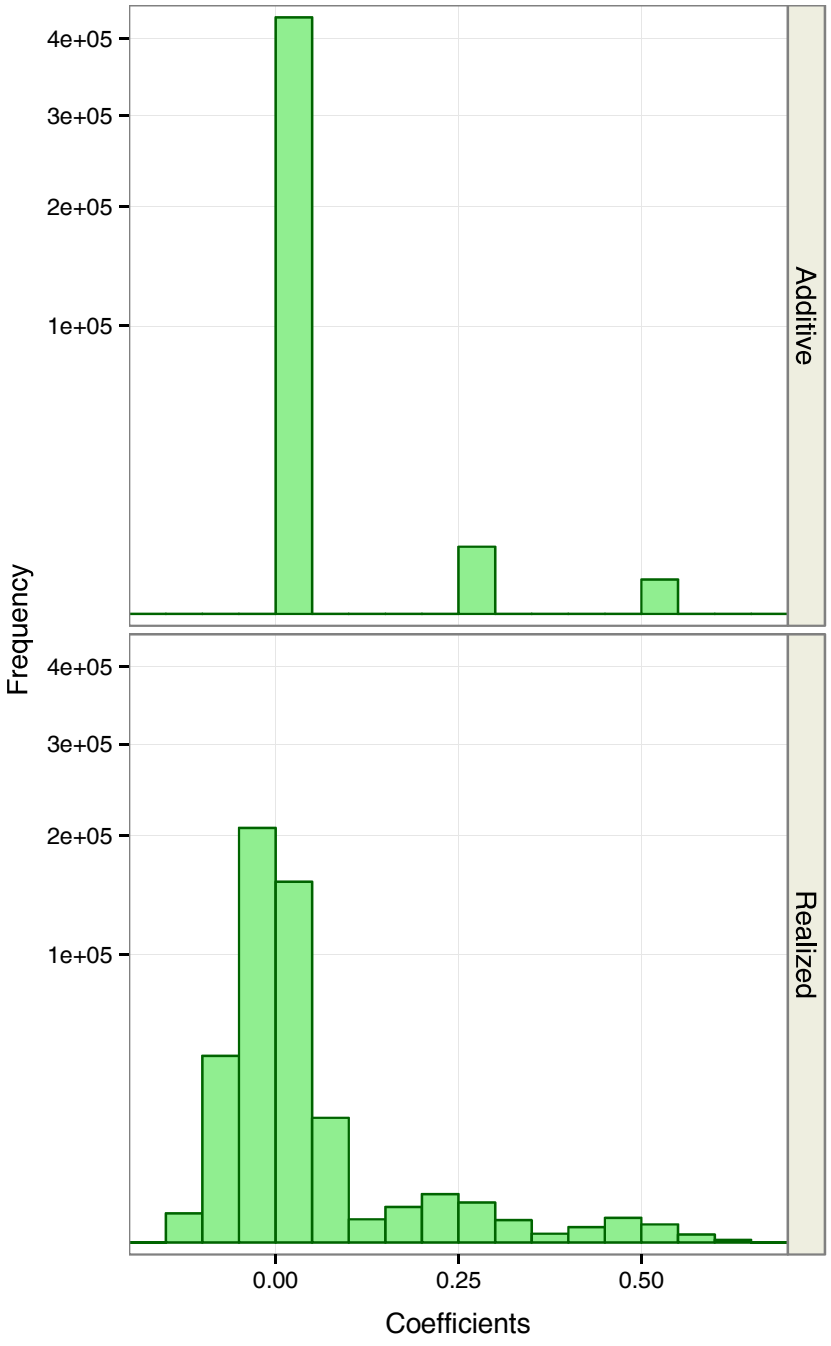


Figure 4: The predictive ability (r) of genomic estimated breeding values (GEBV) for tree height in a validation data set ($n = 48$ sampled from G1 population) using GBLUP, Bayesian ridge and Bayesian LASSO methods (red dots). The blue dots (small dots) represent the training set ($n = 613$ genotypes from G0 and G1).

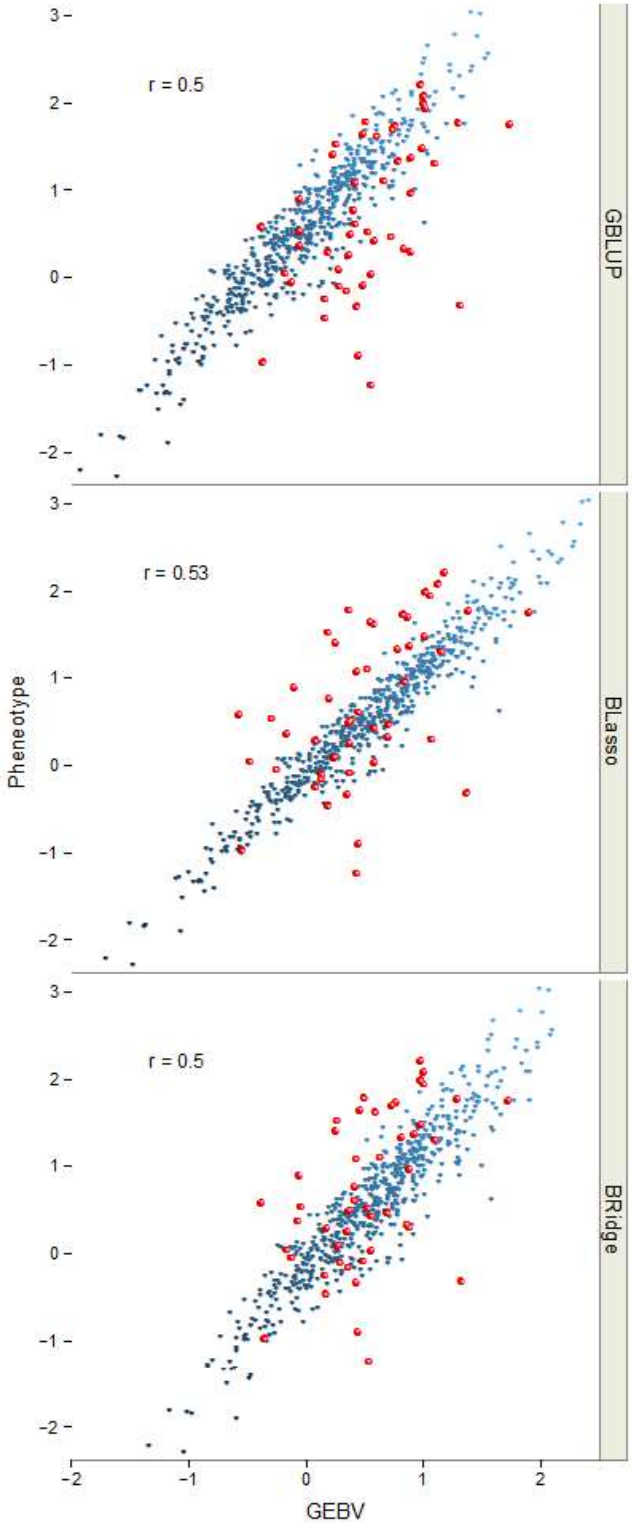


Table 1: Summary statistics of linkage disequilibrium for each linkage group (LG) before correcting for relatedness (r^2) and after correcting for relatedness (r_V^2) between individuals.

LG	Length (cM)	# of markers	r^2			r_V^2		
			Mean	Min	Max	Mean	Min	Max
1	115.0	169	0.014	3.45E-10	1	0.008	2.60E-11	1
2	153.2	180	0.010	2.1E-12	1	0.006	6.10E-12	1
3	140.5	194	0.008	9.32E-13	1	0.005	1.50E-12	1
4	139.5	184	0.010	8.24E-13	1	0.005	5.00E-11	1
5	140.2	164	0.013	3.49E-11	1	0.006	1.30E-13	1
6	114.5	194	0.011	4.48E-11	1	0.007	2.60E-12	1
7	140.5	190	0.010	4.99E-10	1	0.005	3.90E-11	1
8	133.8	152	0.011	9.41E-12	1	0.007	4.70E-13	1
9	111.4	201	0.010	1.34E-11	1	0.005	2.20E-13	1
10	147.4	165	0.009	1.42E-12	1	0.005	6.50E-14	1
11	111.5	197	0.019	4.71E-11	1	0.012	9.80E-12	1
12	144.4	194	0.010	3.79E-12	1	0.006	5.50E-12	1

Table 2: Genetic relationships from pedigree and markers using the shared allele frequency. The average coefficient of relationship is essentially zero (0.0046 from pedigree and -0.0015 from markers) and the average inbreeding coefficient F is close to 0, as expected for a population with only one generation of breeding.

	1+ Inbreeding			**Relationships**		
Methods	Min	Mean	Max	Min	Mean	Max
Pedigree	1.0	1.0	1.0	0	0.0046	0.5
Realized	0.862	0.994	1.139	-0.145	-0.0015	1.085

Table 3. Model evaluation statistics using random sampling of 10% of individuals (48 trees) from G1 population with 100 replications. Model performance statistics for GBLUP, Bayesian ridge regression (BRidge) and Bayesian LASSO (BLasso) for tree diameter, height and stem sweep were presented.

Trait / statistics	GBLUP	BRidge	BLasso
Diameter /	Mean (min-max)	Mean (min-max)	Mean (min-max)
Predictive ability	0.46 (0.14-0.72)	0.47 (0.13-0.72)	0.47 (0.09-0.69)
Rank correlation	0.42 (0.06-0.68)	0.42 (0.06-0.68)	0.41 (0.02-0.67)
Mean squared error	0.42 (0.23-0.65)	0.41 (0.24-0.66)	0.43 (0.25-0.71)
Bias (regression slope)	0.92 (0.27-1.51)	0.91 (0.25-1.50)	0.72 (0.16-1.25)
Mean best 10%	1.08 (0.81-1.31)	1.08 (0.83-1.33)	1.24 (0.93-1.58)
Height /	Mean (min-max)	Mean (min-max)	Mean (min-max)
Predictive ability	0.47 (0.1-0.66)	0.47 (0.11-0.66)	0.46 (0.15-0.67)
Rank correlation	0.44 (0.03-0.67)	0.44 (0.03-0.68)	0.43 (0.05-0.66)
Mean squared error	0.49 (0.31-0.81)	0.49 (0.31-0.81)	0.52 (0.34-0.85)
Bias (regression slope)	0.79 (0.15-1.25)	0.79 (0.16-1.23)	0.65 (0.20-1.11)
Mean best 10%	1.49 (1.15-1.87)	1.49 (1.16-1.88)	1.64 (1.24-2.06)
Stem Sweep /	Mean (min-max)	Mean (min-max)	Mean (min-max)
Predictive ability	0.55 (0.23-0.73)	0.55 (0.22-0.73)	0.54 (0.20-0.73)
Rank correlation	0.55 (0.22-0.76)	0.55 (0.22-0.76)	0.54 (0.18-0.77)
Mean squared error	0.49 (0.27-0.79)	0.49 (0.27-0.79)	0.52 (0.31-0.82)
Bias (regression slope)	0.85 (0.27-1.46)	0.85 (0.27-1.46)	0.73 (0.21-1.21)
Mean best 10%	0.77 (0.39-1.10)	0.77 (0.40-1.10)	0.90 (0.47-1.28)

Table 4: Model evaluation statistics using random sampling across the population with 100 replications. The whole data set ($n = 661$) was split into training ($n = 331$) and validation sets ($n = 330$). Model performance statistics for GBLUP, Bayesian ridge regression (BRidge) and Bayesian LASSO (BLasso) for tree diameter, height and stem sweep are presented.

Trait / statistics	GBLUP	BRidge	BLasso
Diameter /	Mean (min-max)	Mean (min-max)	Mean (min-max)
Predictive ability	0.39 (0.30-0.50)	0.40 (0.31-0.50)	0.38 (0.29-0.48)
Rank correlation	0.37 (0.28-0.49)	0.37 (0.28-0.48)	0.36 (0.26-0.48)
Mean squared error	0.46 (0.39-0.54)	0.46 (0.39-0.54)	0.49 (0.41-0.57)
Bias (regression slope)	0.95 (0.60-1.96)	0.90 (0.61-1.50)	0.63 (0.46-0.82)
Mean best 10%	0.95 (0.65-1.14)	0.97 (0.74-1.15)	1.18 (1.05-1.39)
Height /	Mean (min-max)	Mean (min-max)	Mean (min-max)
Predictive ability	0.47 (0.35-0.55)	0.47 (0.35-0.55)	0.45 (0.34-0.53)
Rank correlation	0.46 (0.34-0.57)	0.46 (0.34-0.56)	0.44 (0.33-0.55)
Mean squared error	0.57 (0.47-0.76)	0.57 (0.47-0.74)	0.60 (0.49-0.77)
Bias (regression slope)	0.98 (0.63-1.74)	0.97 (0.65-1.55)	0.73 (0.48-0.98)
Mean best 10%	1.37 (1.11-1.70)	1.37 (1.14-1.67)	1.56 (1.36-1.82)
Stem Sweep /	Mean (min-max)	Mean (min-max)	Mean (min-max)
Predictive ability	0.49 (0.40-0.57)	0.49 (0.41-0.57)	0.48 (0.36-0.57)
Rank correlation	0.48 (0.41-0.57)	0.48 (0.41-0.57)	0.47 (0.36-0.56)
Mean squared error	0.61 (0.50-0.73)	0.61 (0.50-0.72)	0.63 (0.52-0.75)
Bias (regression slope)	0.97 (0.53-1.64)	0.97 (0.62-1.50)	0.77 (0.55-1.03)
Mean best 10%	0.69 (0.41-1.00)	0.68 (0.47-0.85)	0.86 (0.72-1.05)

Genomic selection in maritime pine

Supplemental Figures and Tables

Fikret Isik [1,2] , Jérôme Bartholomé [1,3],

Alfredo Farjat [2,4], Emilie Chancerel [1,3], Annie Raffin [1,3], Leopoldo Sanchez [5],

Christophe Plomion [1,3], Laurent Bouffier [1,3*]

1/ INRA, UMR1202, BIOGECO, Cestas F-33610, France
2/ Department of Forestry and Environmental Resources, North Carolina State University, Raleigh, NC, USA
3/ Univ. Bordeaux, BIOGECO, UMR1202, Talence F-33170, France
4/ Department of Statistics, North Carolina State University, Raleigh, NC, USA
5/ INRA, Orleans, France

* For correspondence
Laurent Bouffier,
Address: INRA, UMR1202, BIOGECO, Cestas F-33610, France
Email: bouffier@pierroton.inra.fr
The following Supporting Information is available for this article:

Figure S1: Pedigree of maritime pine breeding population
Figure S2: Genotype calls in the population
Figure S3: High density genetic map of maritime pine and LD on each chromosome
Figure S4: LD heat map for all linkage groups before (r^2, A) and after (r_v^2, B) correction of relatedness. See separate files
Figure S5: Relationship between the two estimates of linkage disequilibrium (r^2 and r^2_v).
Figure S6: Inbreeding coefficients of 661 individuals derived from 2500 SNP markers.
Figure S7A: Evolution of the Gelman and Rubin's shrink factor. **Figure S7**B: Bayesian convergence monitoring plots for the residual variance.
Figure S8: Priors and estimated scaled posterior densities for the residual variance σ_e^2
Figure S9: Comparisons of three statistical models for GEBV
Table S1: Performance of statistical models on the whole dataset.
Table S2: Cross validation using random sampling of 10% from G1 population for de-regressed breeding values.

Figure S1: Pedigree of maritime pine breeding population. Pedigree Viewer software (http://www-personal.une.edu.au/~bkinghor/pedigree.htm) was used to visualize the relatedness between two generations. The population is composed of two tiers with 184 individuals in G0 and 477 individuals in G1. Black lines indicate link between female parents and their progeny and blue lines indicate link between male parents and their progeny. Average inbreeding and maximum inbreeding coefficients were zero in the population because of unrelated founders and one selection cycle. The maximum maternal and paternal family sizes were 13 and 25, respectively.

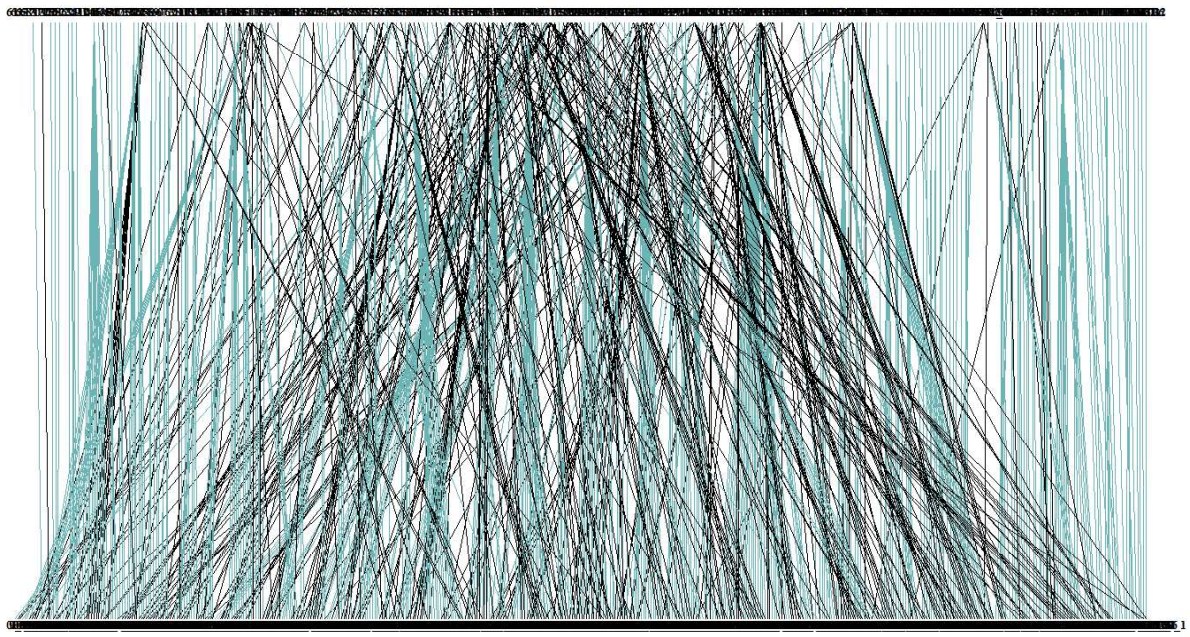


Figure S2: Genotype calls in the population. The frequency of genotype calls ranged from 0.17 (AA) to 0.025 (AT/TA). A high proportion of genotype calls (61%) were homozygous (GG, CC, AA and TT).

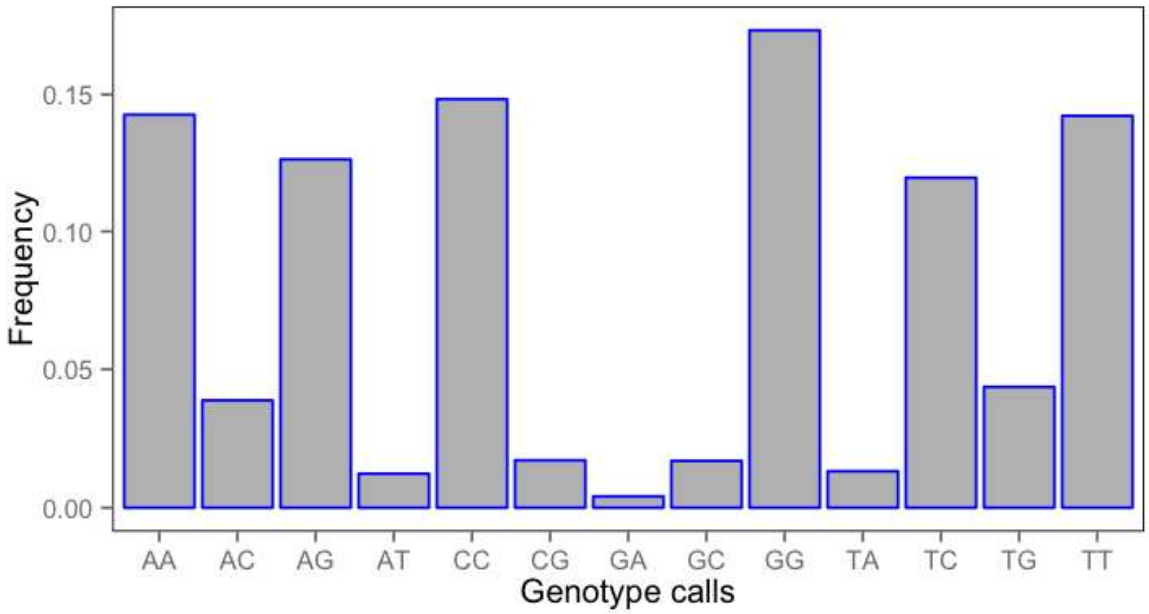


Figure S3: High density genetic map of maritime pine and LD on each chromosome. Positions of the 2,183 mapped markers on the 12 chromosomes are illustrated. The number of markers is given at the bottom of each chromosome.

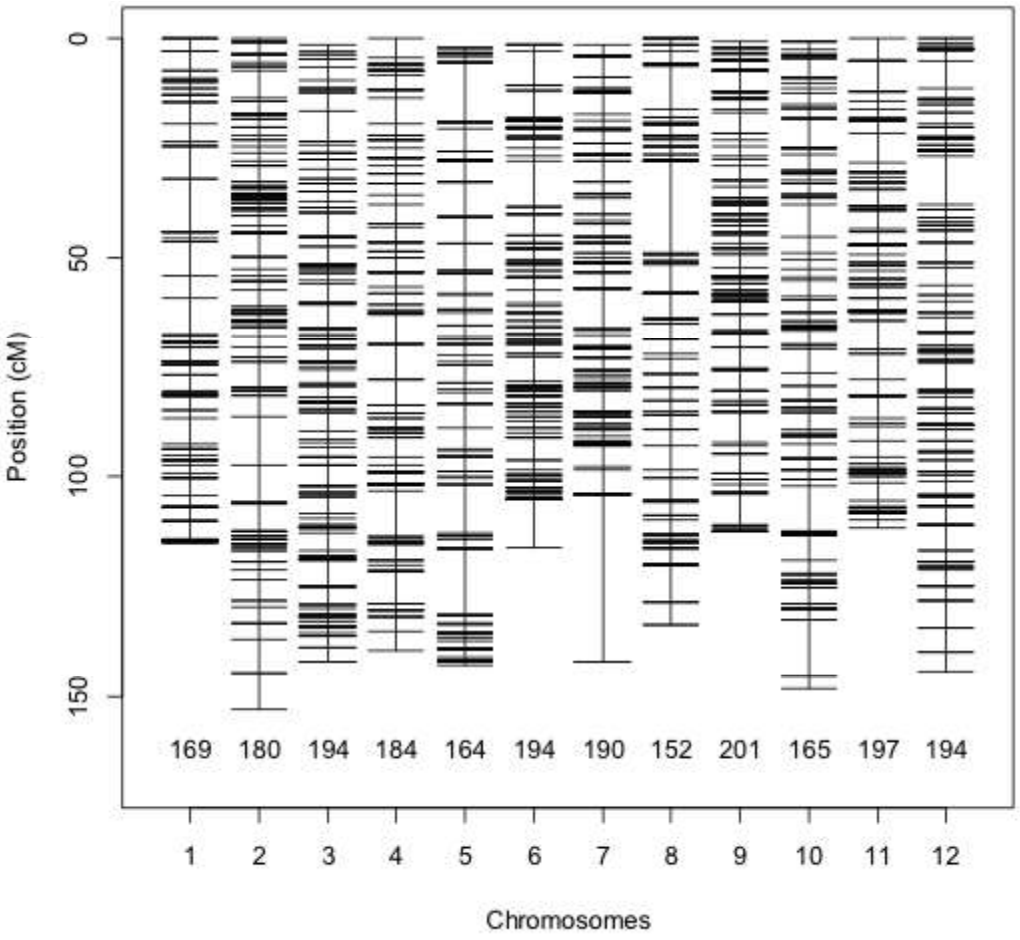


Figure S4: See separate files

Figure S5: Relationship of two estimates of LD (r^2 and r^2_v) between all pairs of markers (≈199,000 pairwise estimates) on linkage groups. Different classes of inter-marker distances are represented. The data showed that regular LD estimates are biased upward because of relatedness between individuals.

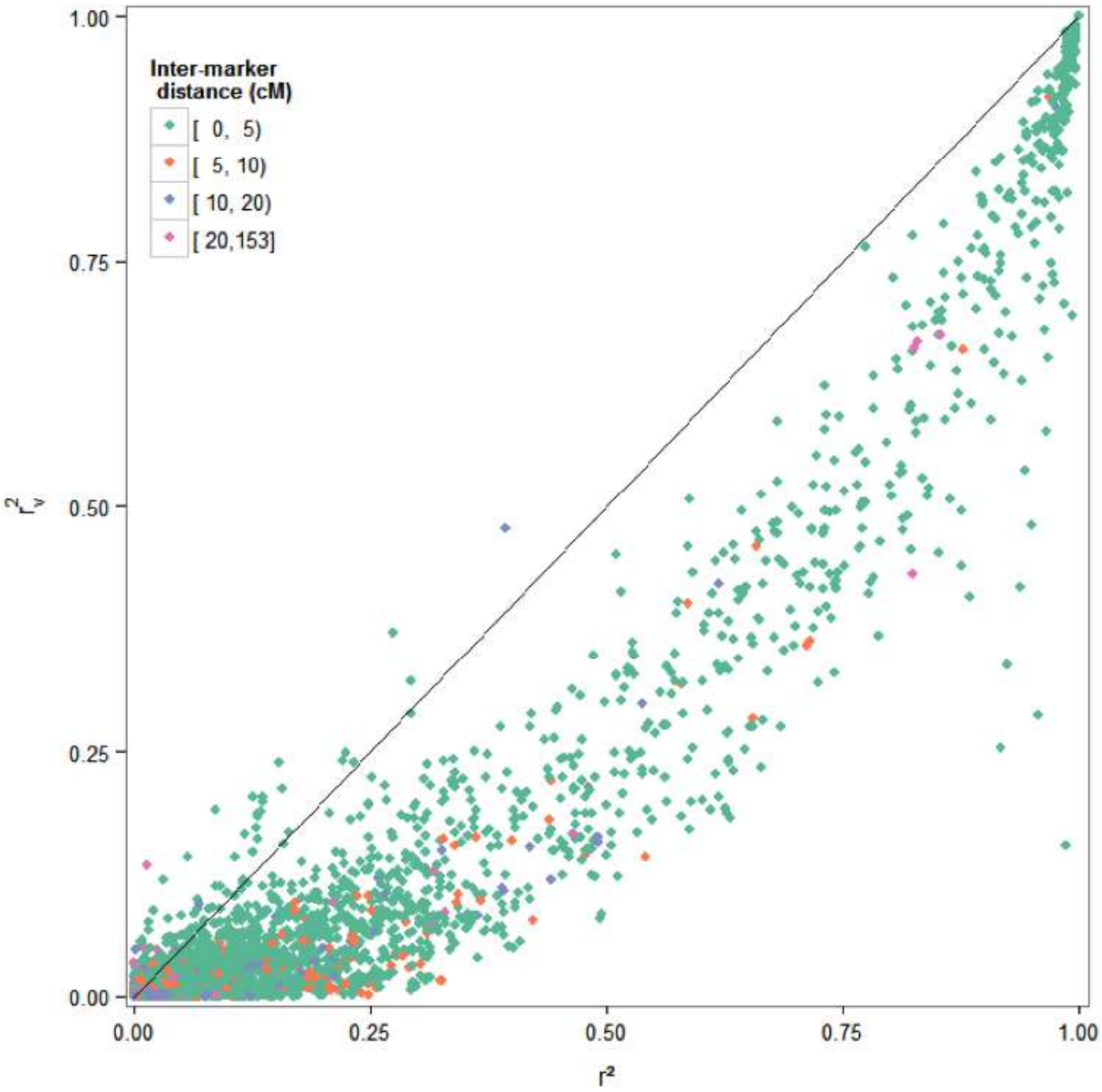


Figure S6: Inbreeding coefficients of 661 individuals derived from 2500 SNP markers. For a non-bred population the expected inbreeding coefficient is 1 ± 0.2 Std dev. Too high and too low inbreeding values indicate genotyping mistakes and/or errors in pedigree [78]. The inbreeding estimates in our data are within the expected boundaries.

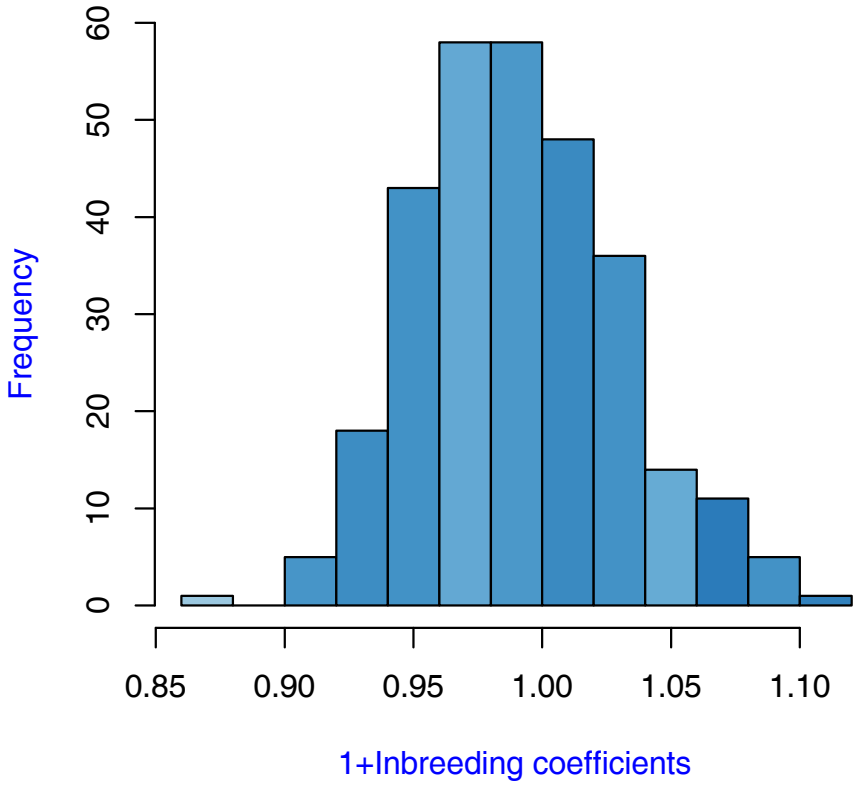


Figure S7A: Evolution of the Gelman and Rubin's shrink factor as the number of iterations increases for five MCMC chains of the residual variance for height as the response variable. The Bayesian ridge (Right) and Bayesian LASSO (Left) models converge after 10000 iterations.

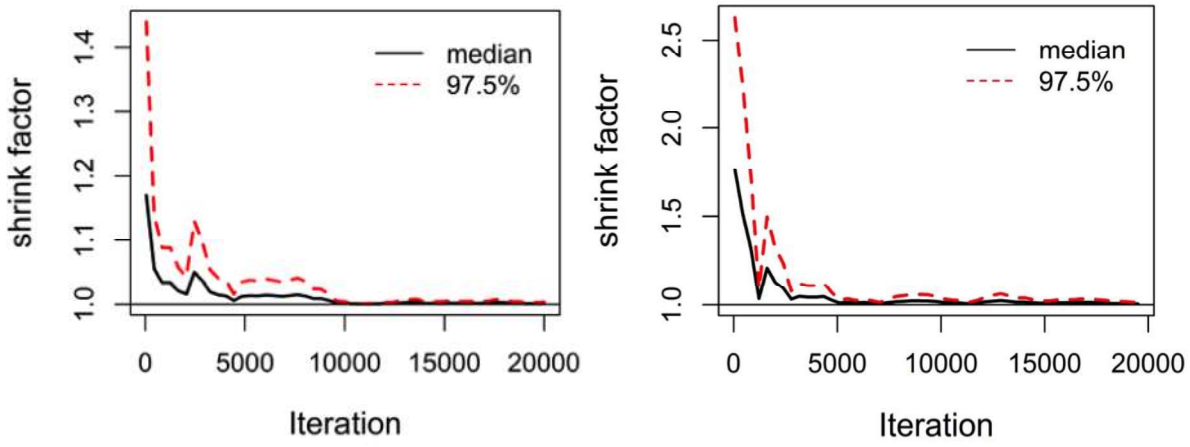


Figure S7B: Bayesian ridge (upper panel) and Bayesian LASSO (lower panel) convergence monitoring plots for the residual variance. The trace plots and autocorrelation functions used 40,000 samples. The burn-in period was set to 10,000 based on the Gelman and Rubin's convergence diagnostic with five parallel chains.

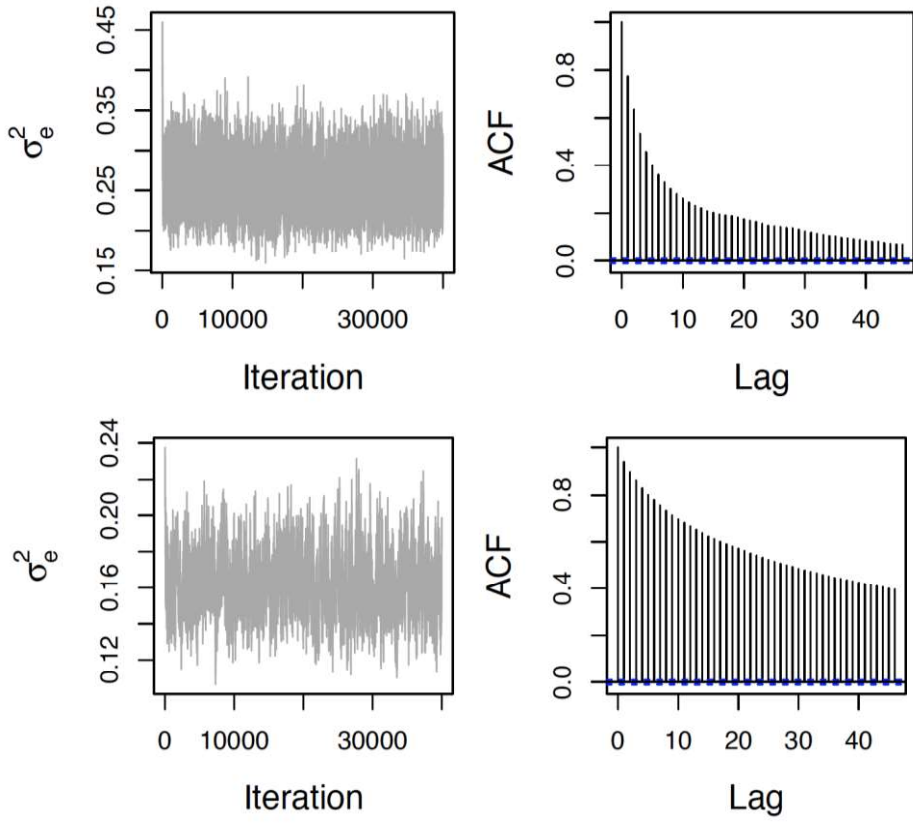


Figure S8: Priors (dashed lines) and estimated scaled posterior densities (solid lines) for the residual variance σ_e^2 assuming three informative priors for height trait for Bayesian ridge (A) and Bayesian LASSO (B) regression models. The prior densities were set assuming that the proportion of phenotypic variance attributed to the residuals were 80% (red), 60% (green), and 40% (blue).

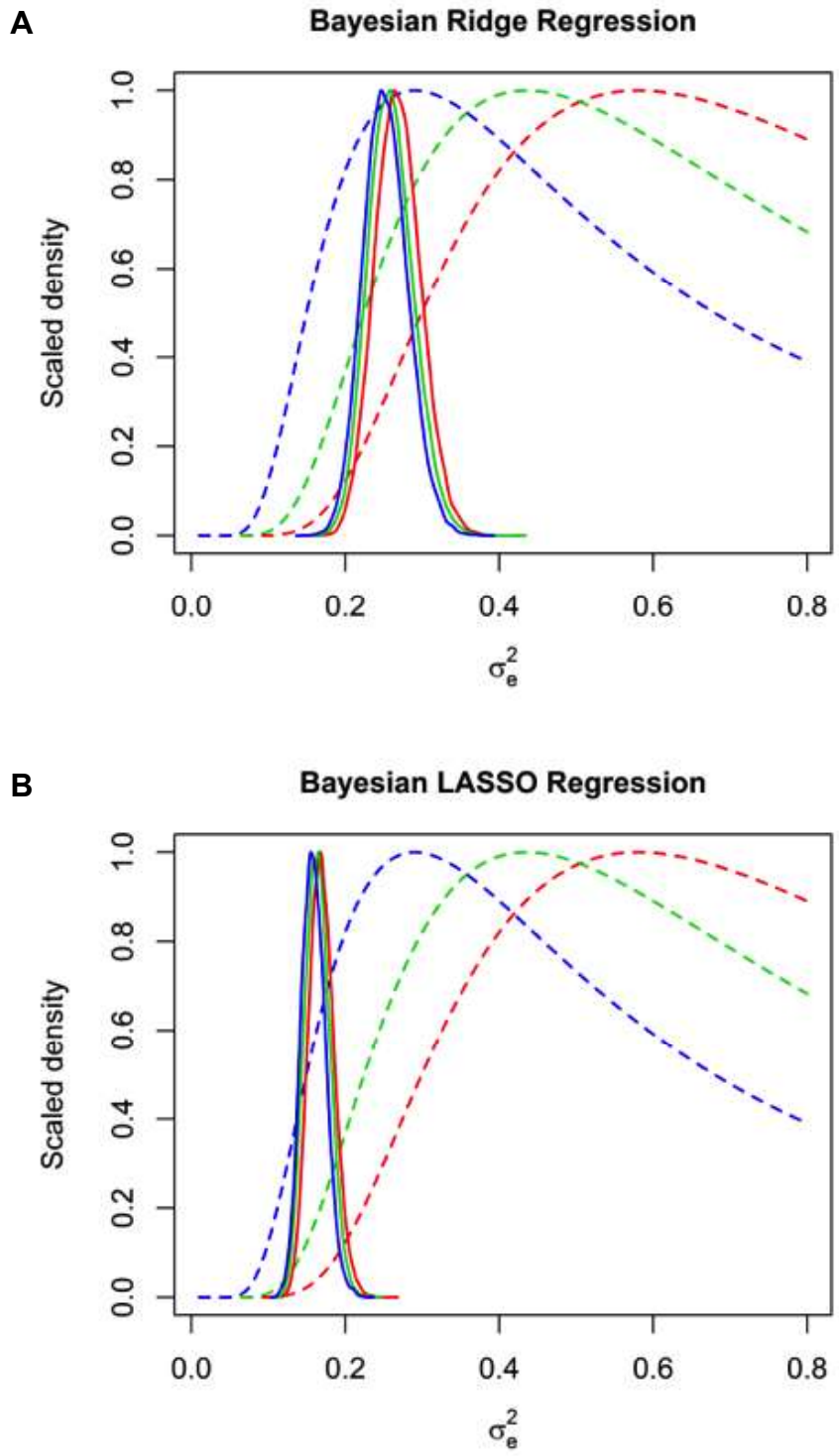
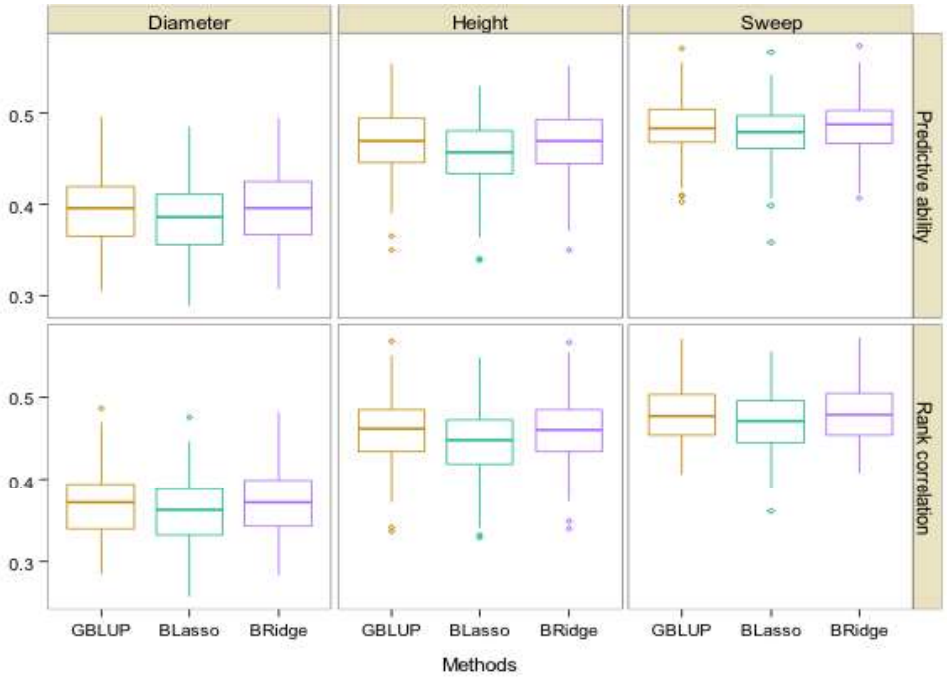


Figure S9: Comparisons of three statistical models for genomic estimated breeding values (GEBV). The box plots show the distribution of 100 predictive ability and rank correlations for three traits. The vertical lines in the middle of the boxes are the medians. The minimum and maximum estimates are shown as circles below and above boxes, respectively.

A) Sampling 50% of the whole population



B) Sampling 10% of the G1 population

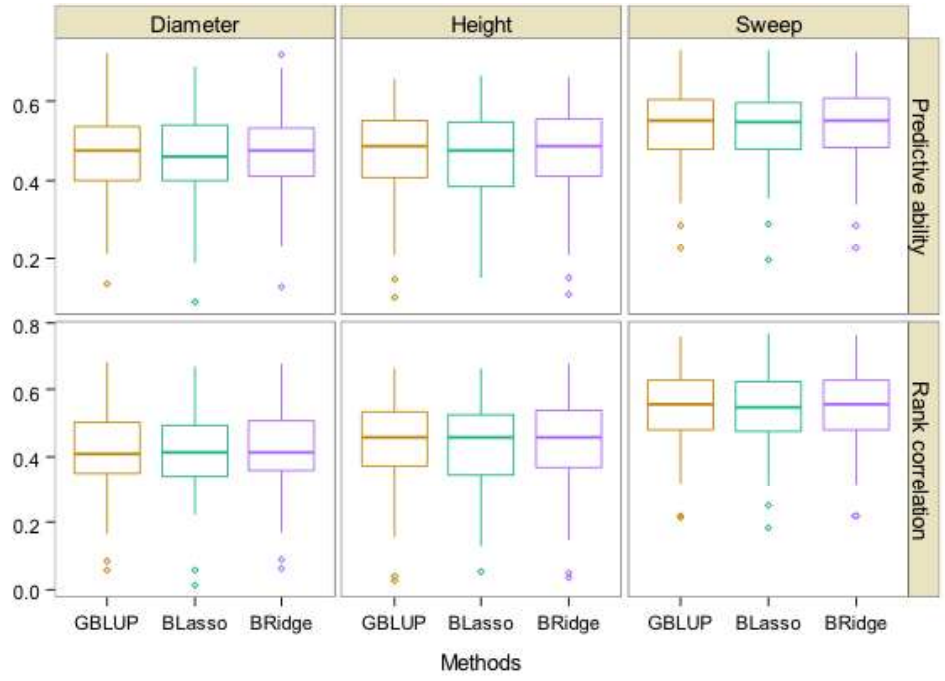


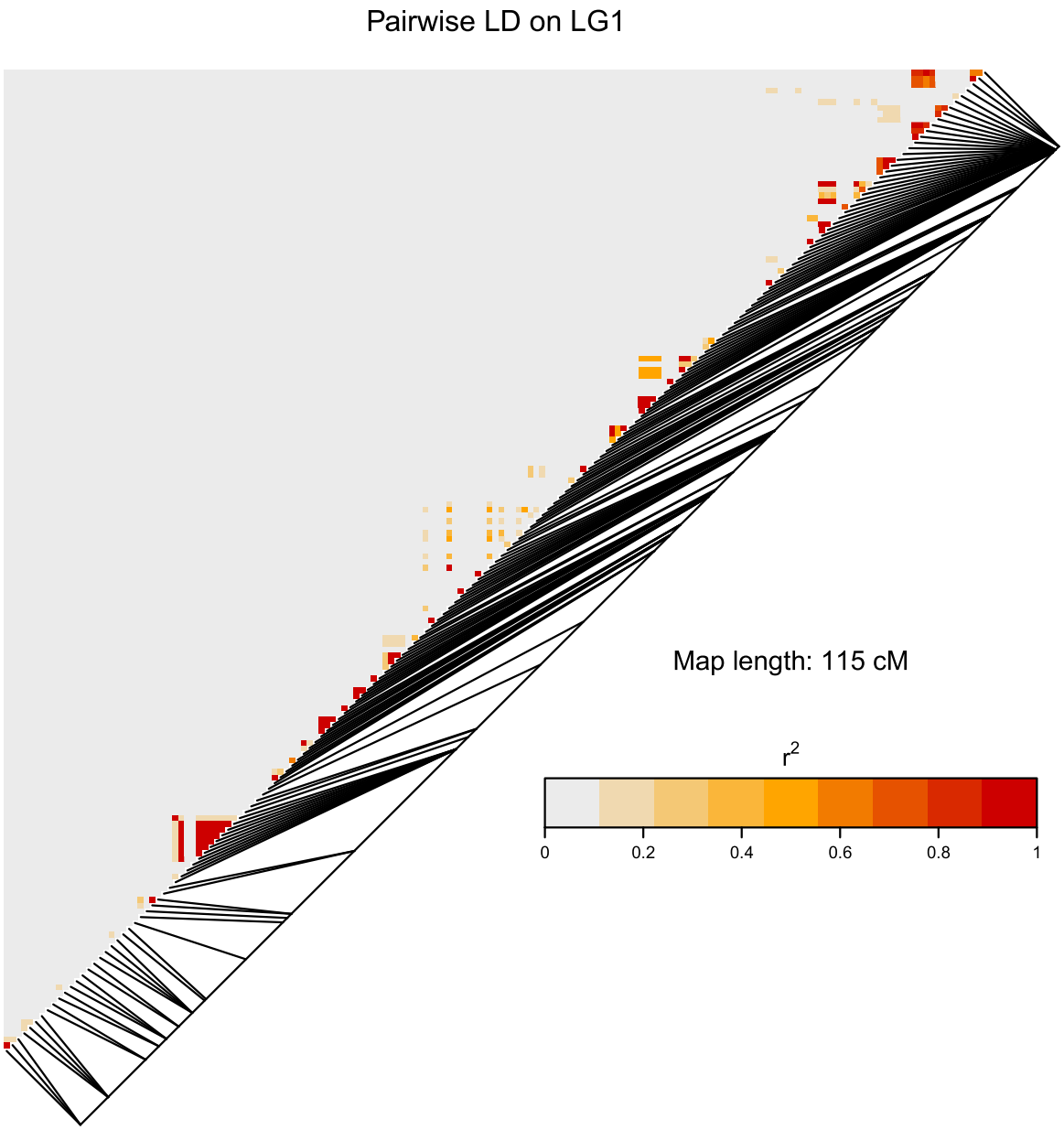
Table S1: Performance of statistical models Bayesian ridge (BRidge) and Bayesian LASSO (BLasso) on the whole dataset (no sampling for validation). BLasso model was clearly superior when all the performance statistics were taken into account.

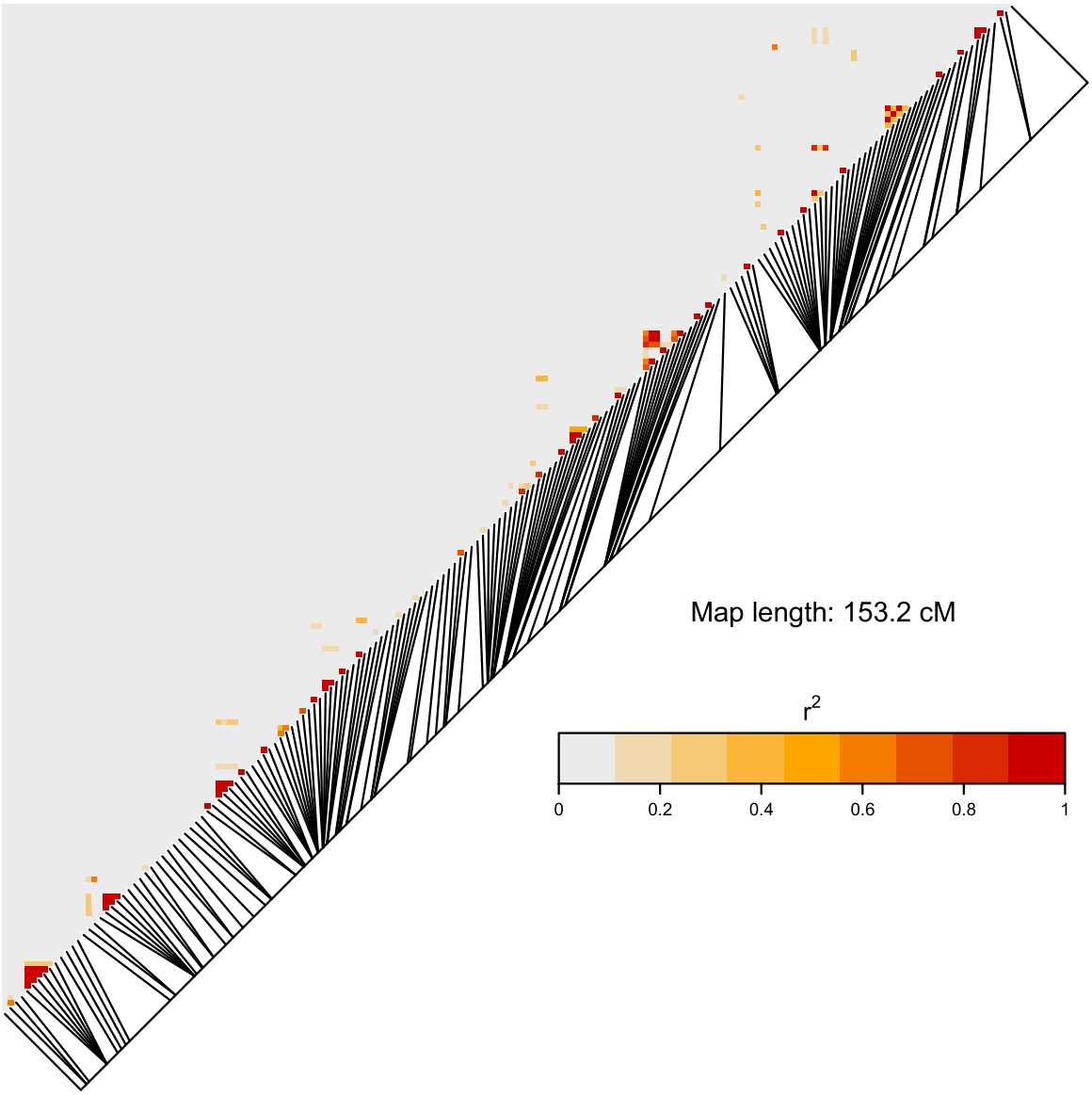
Trait	Statistics	GBLUP	BRidge	BLasso
Diameter	Predictive ability	0.92	0.92	0.97
	Rank correlation	0.91	0.91	0.96
	Mean squared error	0.25	0.12	0.05
	Bias (regression slope)	1.46	1.45	1.23
	Mean best 10%	1.08	1.16	1.34
Height	Predictive ability	0.93	0.94	0.97
	Rank correlation	0.93	0.93	0.96
	Mean squared error	0.47	0.13	0.06
	Bias (regression slope)	1.35	1.34	1.19
	Mean best 10%	1.53	1.61	1.76
Sweep	Predictive ability	0.94	0.94	0.97
	Rank correlation	0.94	0.94	0.96
	Mean squared error	0.14	0.13	0.07
	Bias (regression slope)	1.32	1.32	1.19
	Mean best 10%	1.05	1.08	1.25

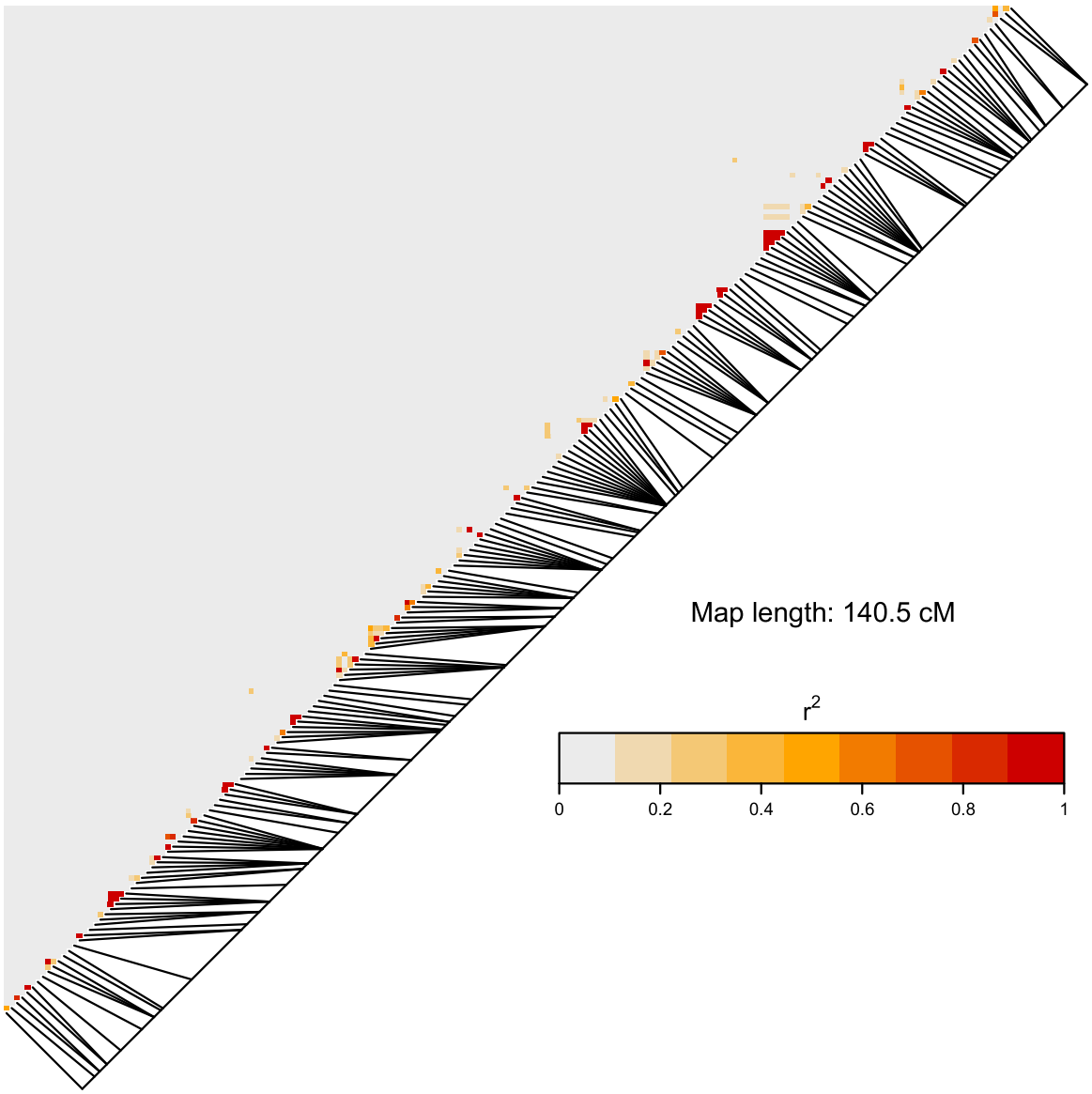
Table S2: Cross validation using random sampling (without replacement) of 10% from G1 population. Model performance statistics for GBLUP, Bayesian ridge regression (BRidge) and Bayesian LASSO (BLasso) for de-regressed tree diameter, height and stem sweepness were presented

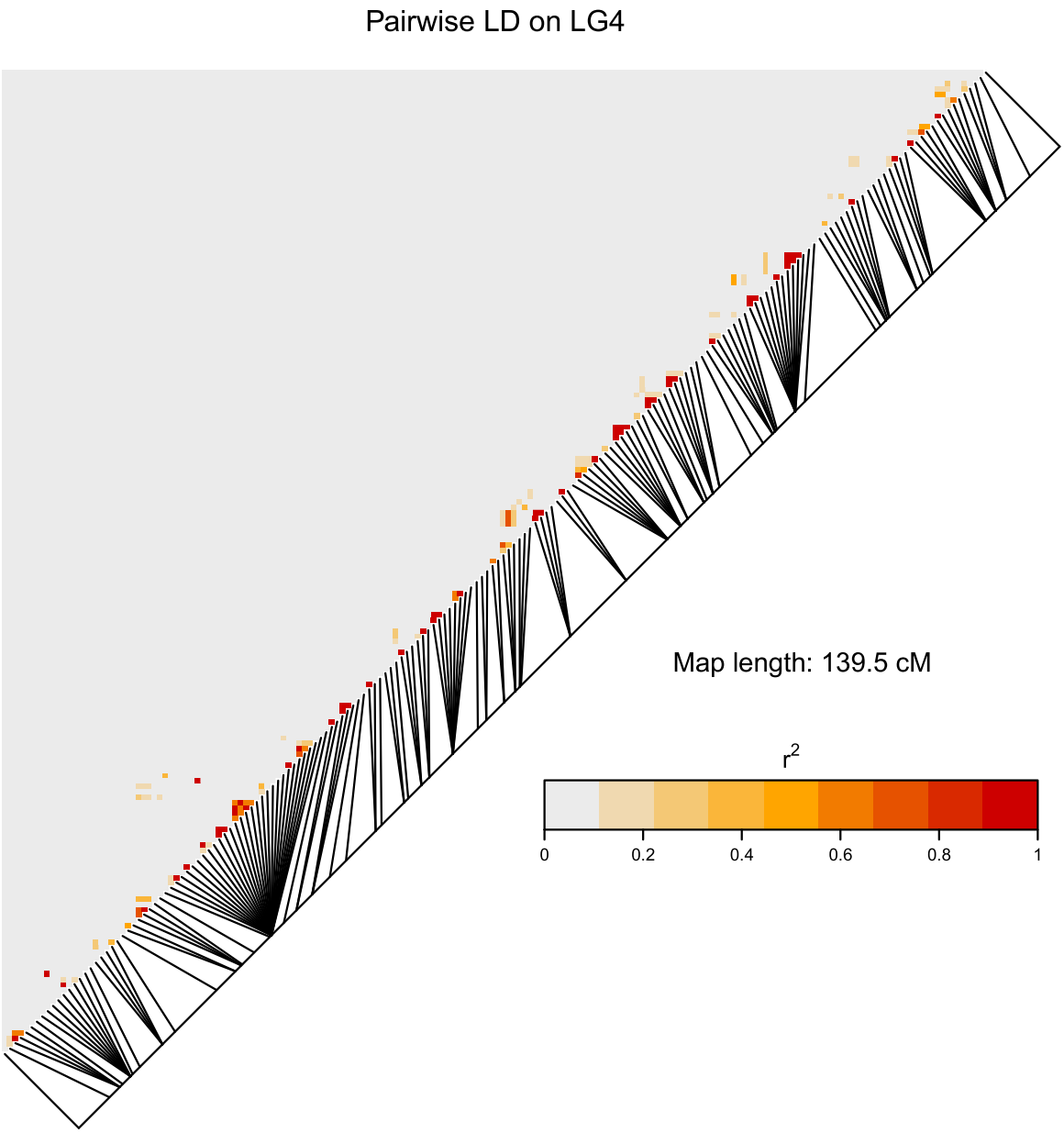
Trait / statistics	GBLUP	BRidge	BLasso
Diameter /	Mean (min-max)	Mean (min-max)	Mean (min-max)
Predictive ability	0.43 (0.04-0.72)	0.43 (0.04-0.72)	0.42 (0.01-0.69)
Rank correlation	0.46 (0.09-0.73)	0.47 (0.09-0.74)	0.46 (0.04-0.70)
Mean squared error	0.53 (0.29-0.89)	0.52 (0.29-0.83)	0.54 (0.32-0.84)
Bias (regression slope)	0.98 (0.17-2.03)	0.98 (0.18-1.97)	0.75 (0.05-1.45)
Mean best 10%	1.14 (0.89-1.42)	1.14 (0.90-1.43)	1.31 (1.01-1.72)
Height /	Mean (min-max)	Mean (min-max)	Mean (min-max)
Predictive ability	0.48 (0.10-0.71)	0.49 (0.11-0.72)	0.47 (0.00-0.70)
Rank correlation	0.50 (0.07-0.71)	0.51 (0.07-0.71)	0.50 (0.02-0.69)
Mean squared error	0.59 (0.37-0.94)	0.59 (0.37-0.94)	0.62 (0.37-1.00)
Bias (regression slope)	0.90 (0.13-1.45)	0.90 (0.14-1.45)	0.74 (0.03-1.18)
Mean best 10%	1.60 (1.30-1.91)	1.59 (1.30-1.90)	1.75 (1.38-2.10)
Stem Sweep /	Mean (min-max)	Mean (min-max)	Mean (min-max)
Predictive ability	0.55 (0.24-0.74)	0.55 (0.24-0.74)	0.55 (0.26-0.76)
Rank correlation	0.55 (0.26-0.74)	0.55 (0.27-0.74)	0.55 (0.29-0.71)
Mean squared error	0.62 (0.33-0.91)	0.61 (0.33-0.91)	0.64 (0.34-0.95)
Bias (regression slope)	0.84 (0.30-1.18)	0.83 (0.30-1.18)	0.77 (0.34-1.20)
Mean best 10%	0.90 (0.34-1.48)	0.90 (0.35-1.47)	0.96 (0.40-1.34)

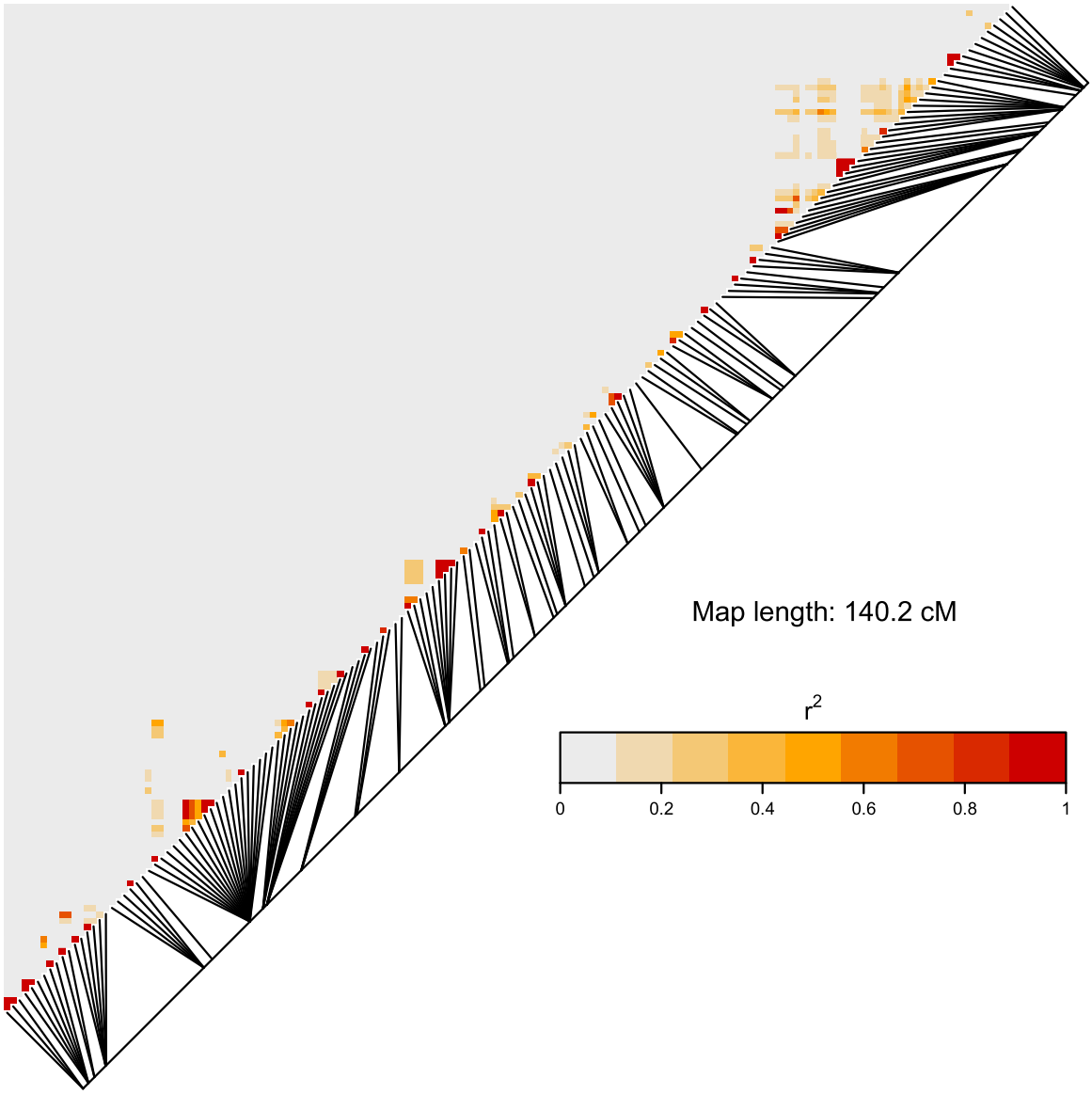
supplemental S4A

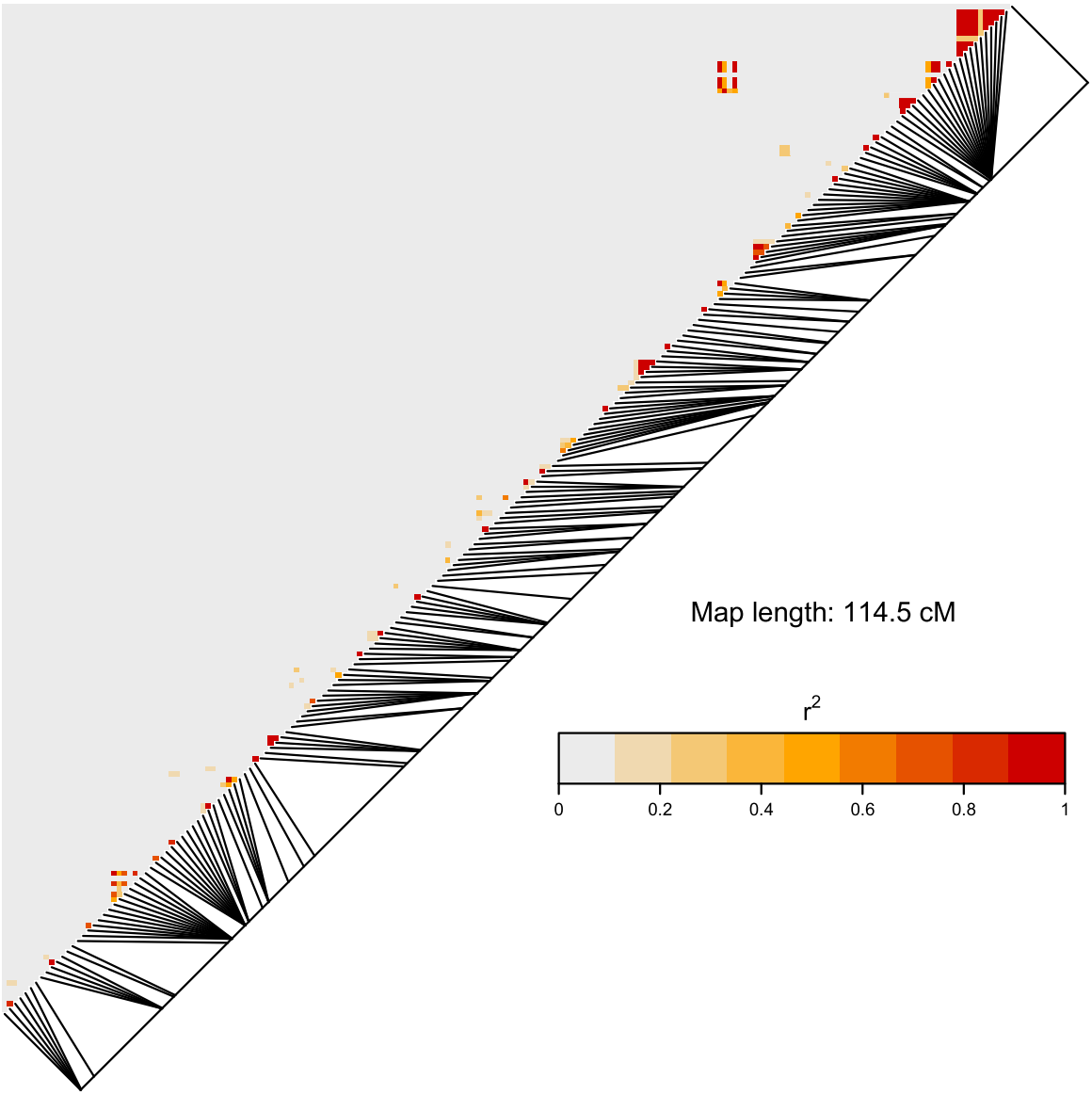


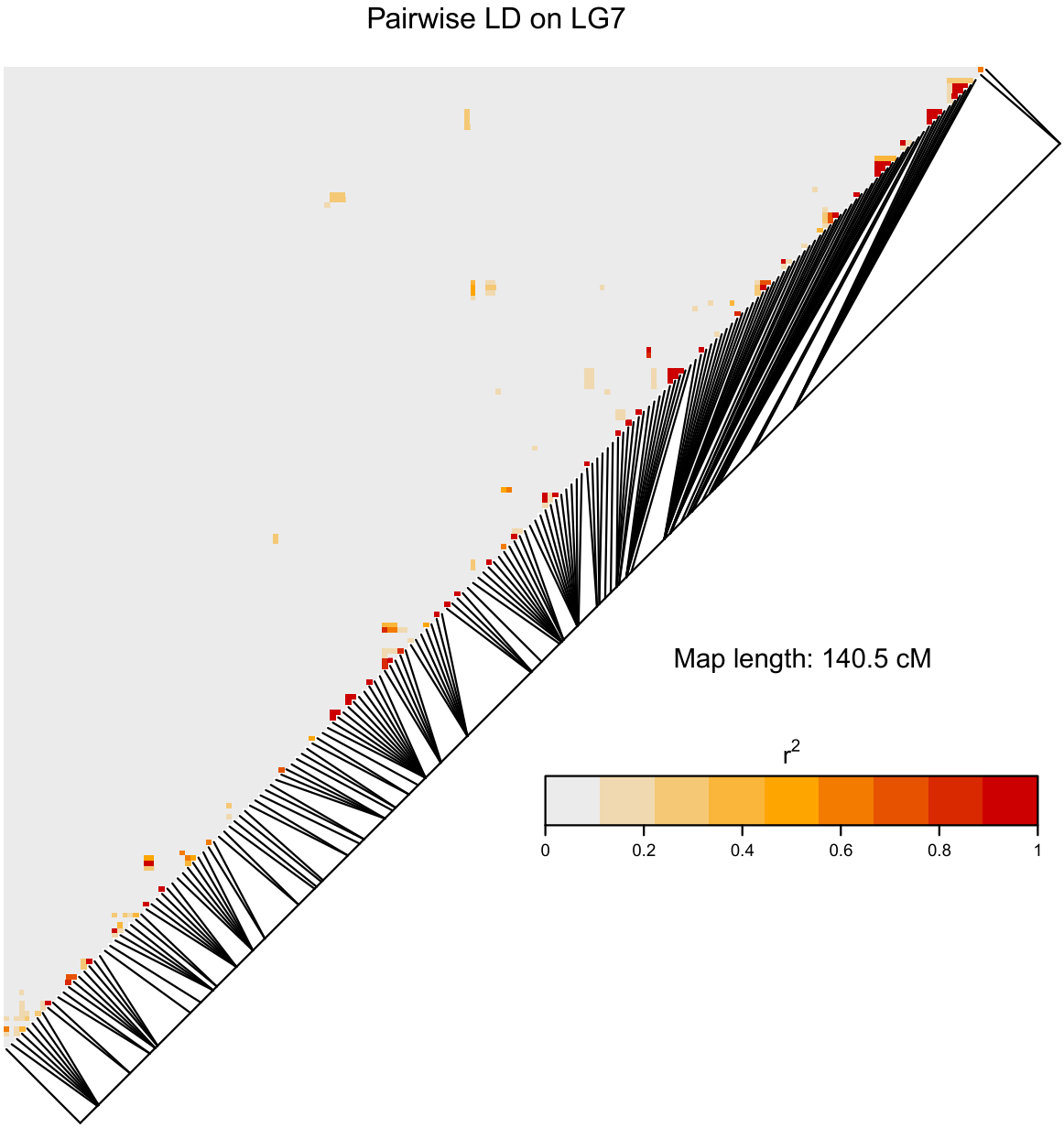

Pairwise LD on LG2
Map length: 153.2 cM
r2
0
0.2
0.4
0.6
0.8
1

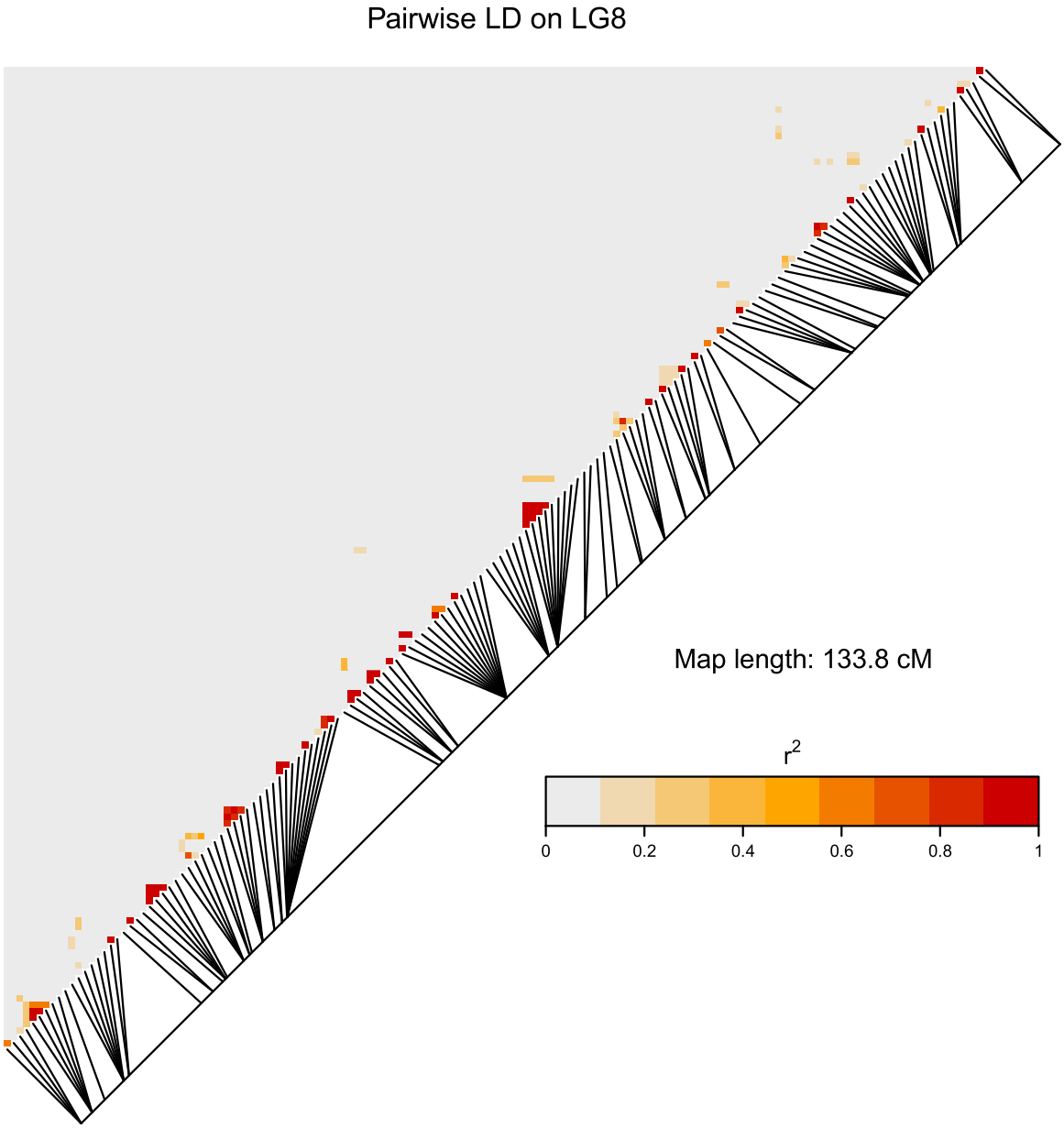

Pairwise LD on LG3
Map length: 140.5 cM
r2
0
0.2
0.4
0.6
0.8
1

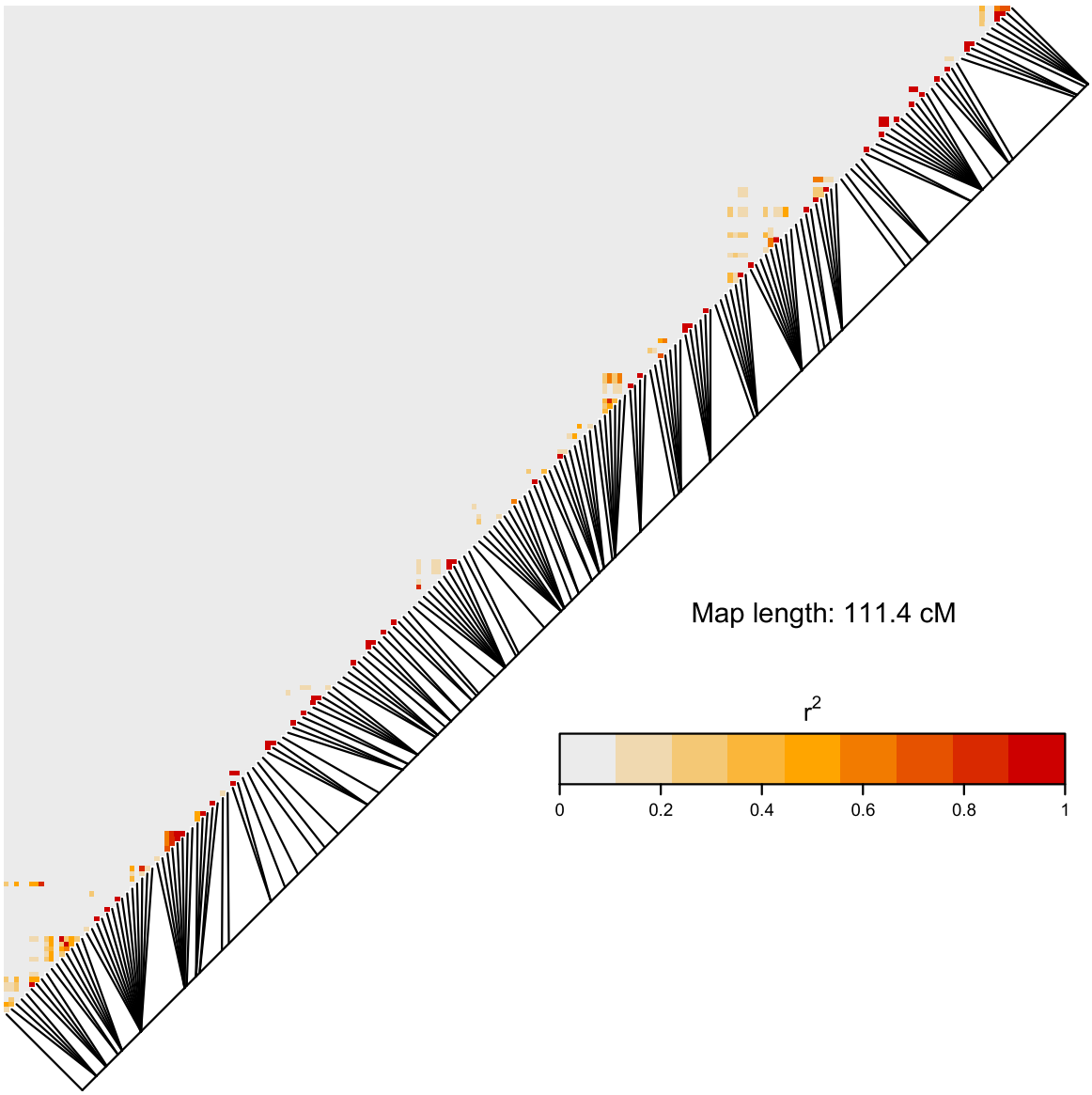

Pairwise LD on LG4
Map length: 139.5 cM
r2
0
0.2
0.4
0.6
0.8
1

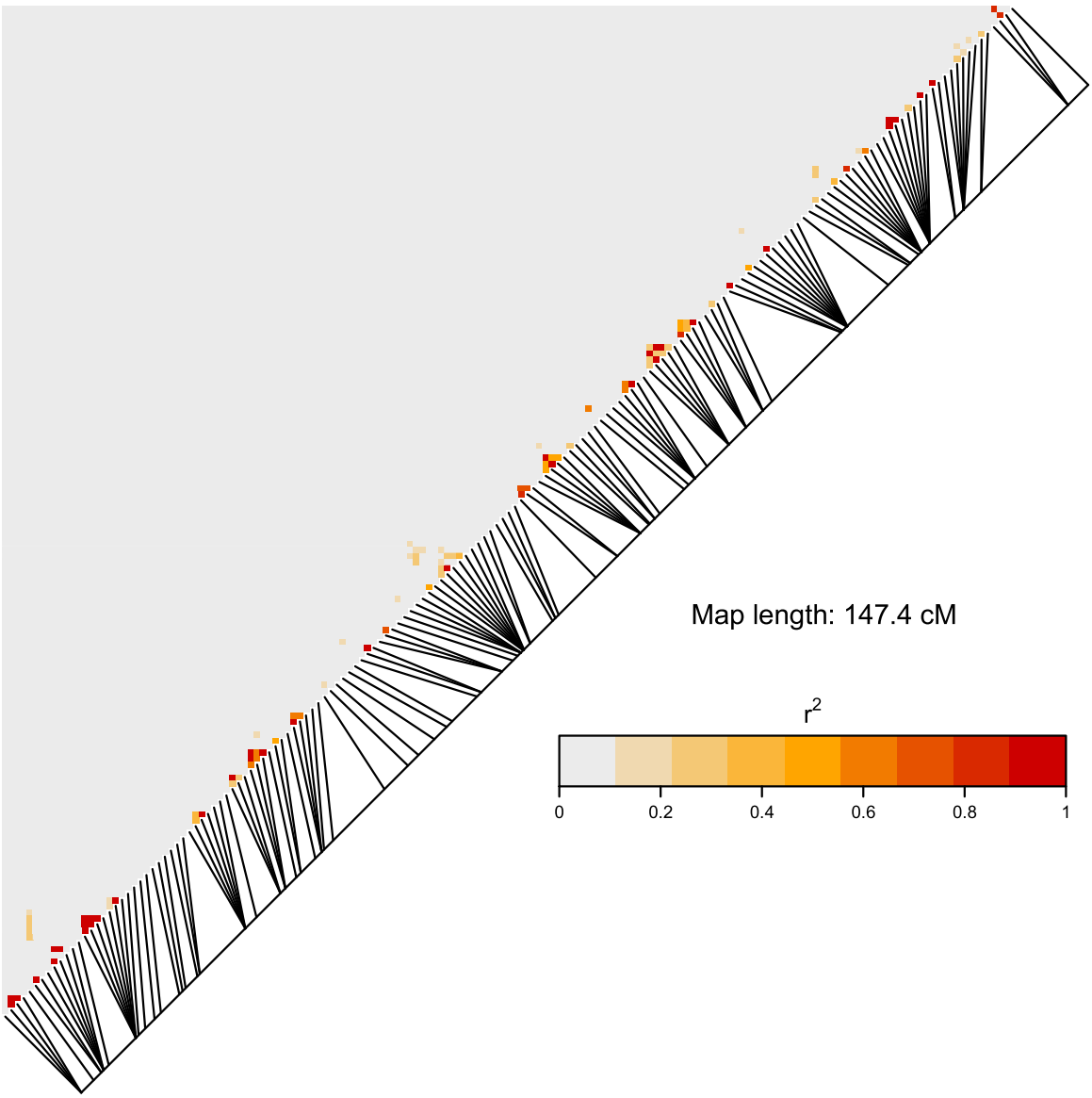

Pairwise LD on LG5
Map length: 140.2 cM
r2
0
0.2
0.4
0.6
0.8
1

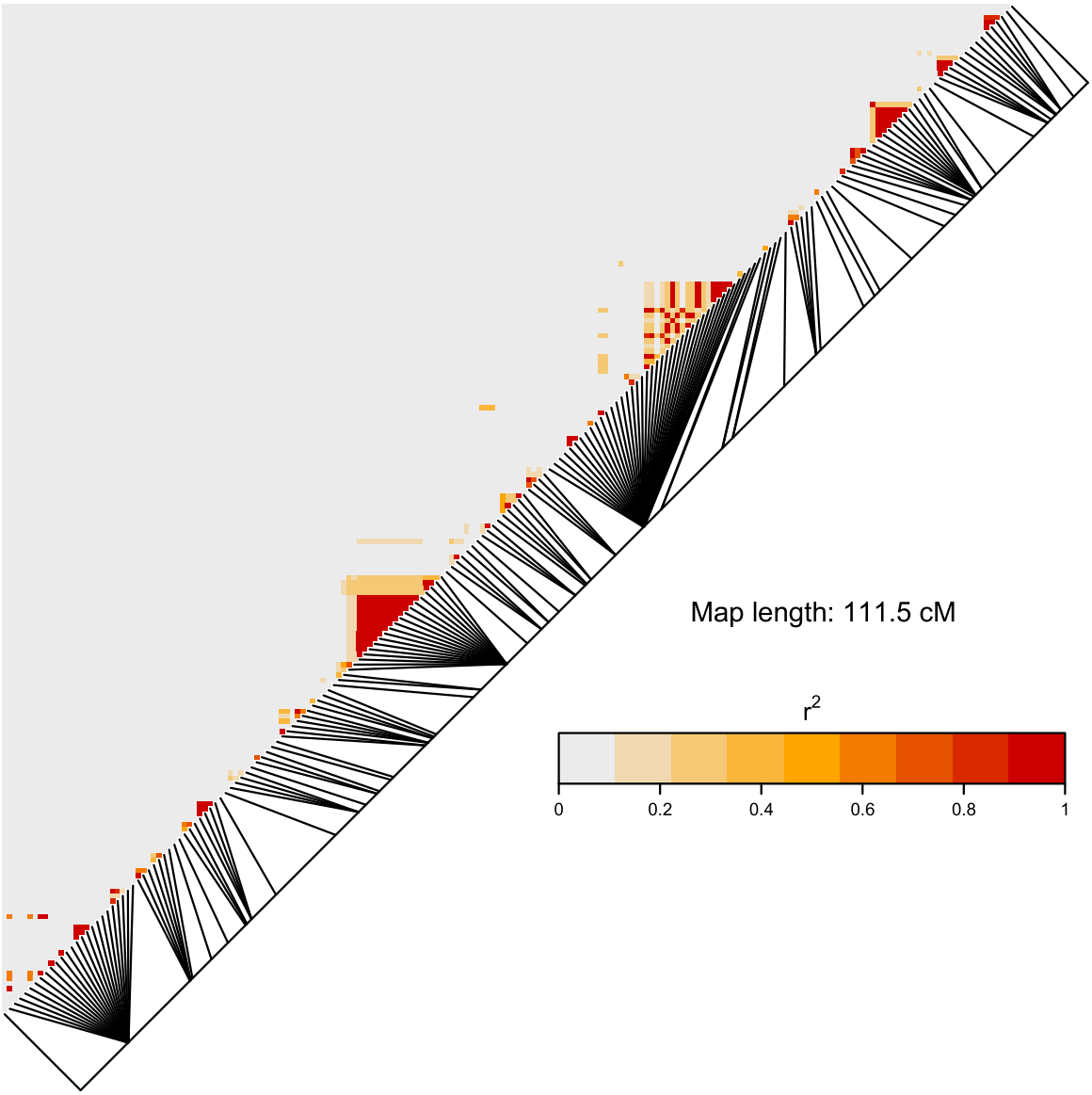

Pairwise LD on LG6
Map length: 114.5 cM
r2
0
0.2
0.4
0.6
0.8
1

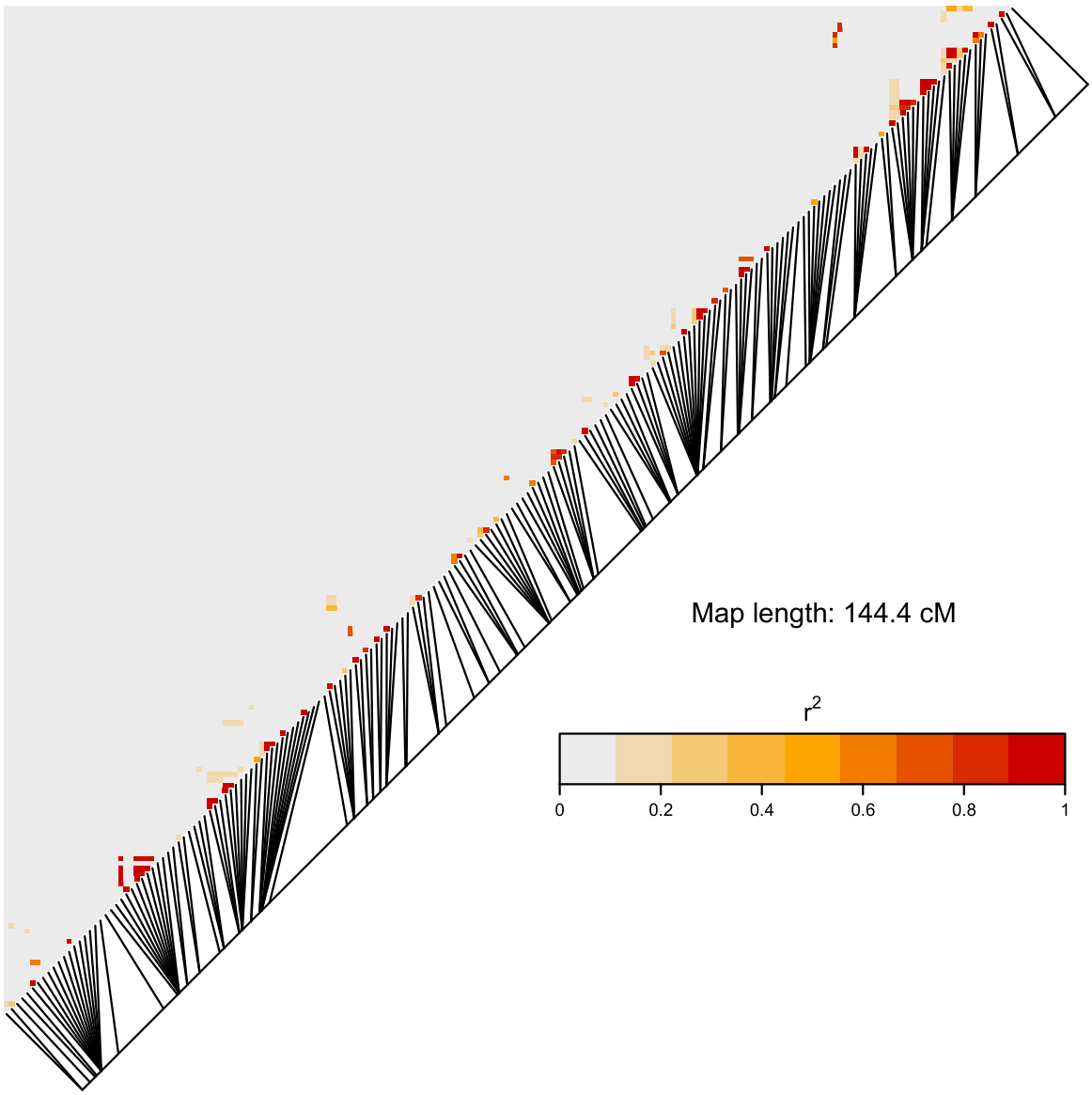

Pairwise LD on LG7
Map length: 140.5 cM
r2
0
0.2
0.4
0.6
0.8
1


Pairwise LD on LG8
Map length: 133.8 cM
r2
0
0.2
0.4
0.6
0.8
1

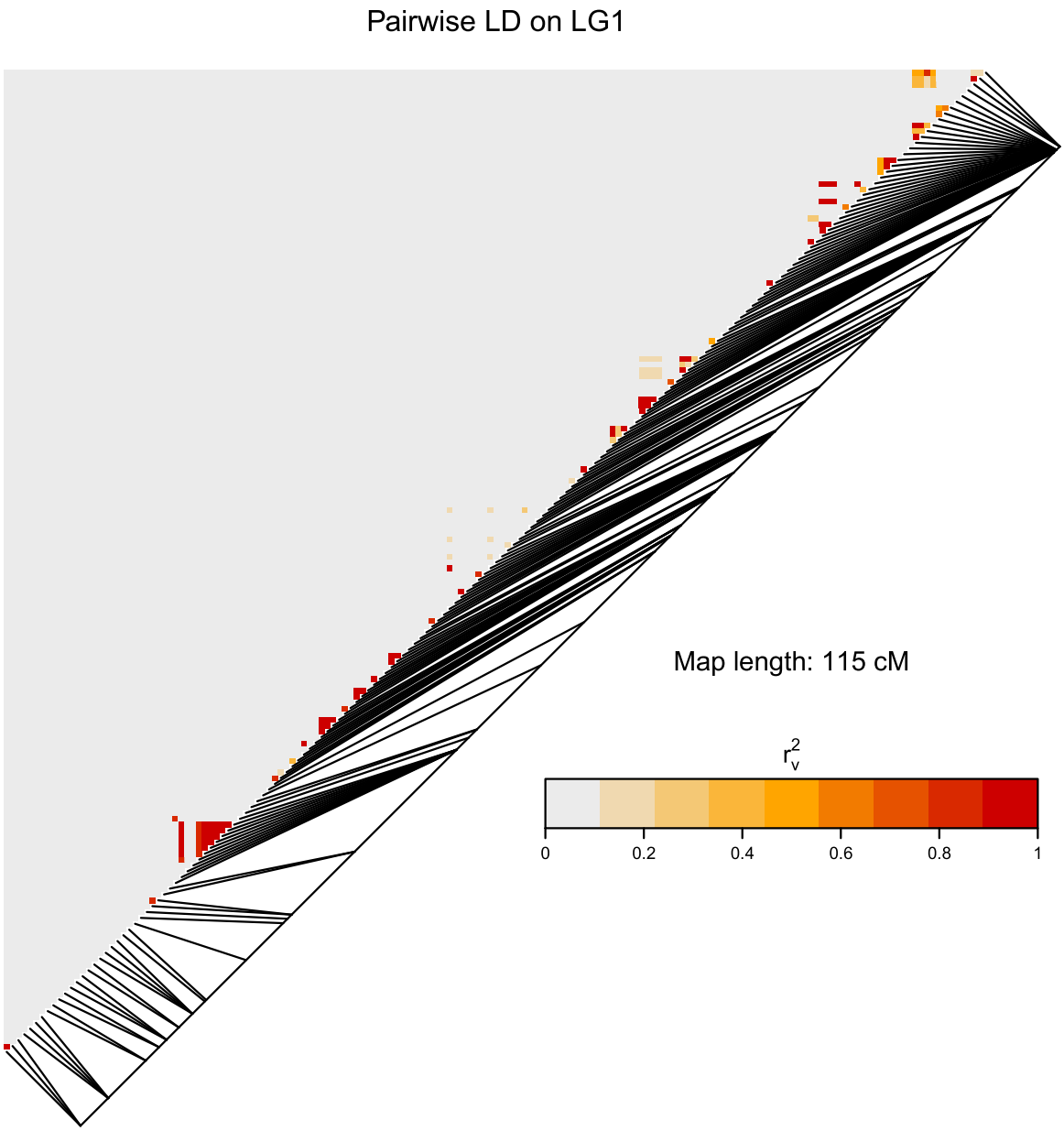

Pairwise LD on LG9
Map length: 111.4 cM
r2
0
0.2
0.4
0.6
0.8
1

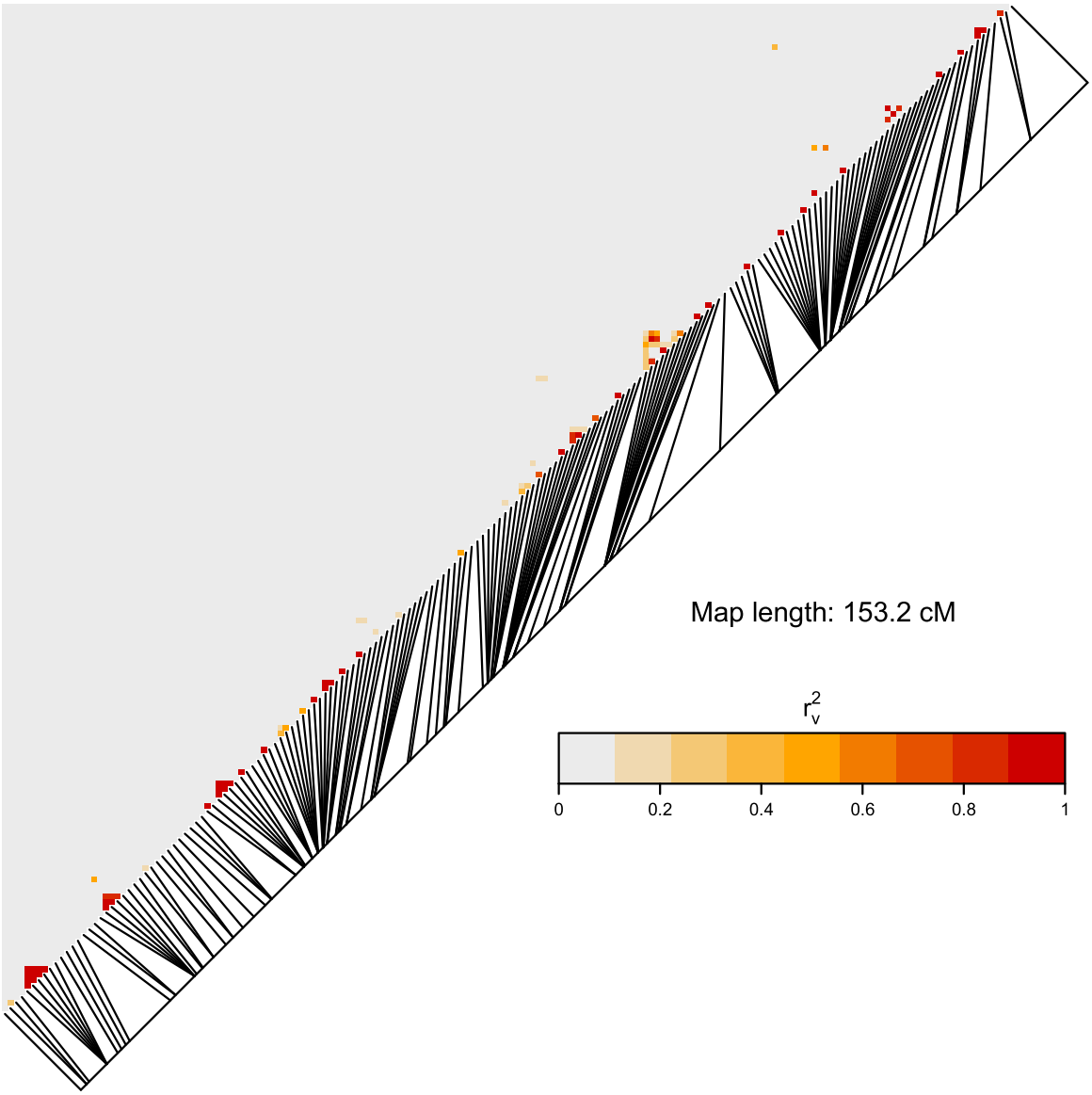

Pairwise LD on LG10
Map length: 147.4 cM
r2
0
0.2
0.4
0.6
0.8
1


Pairwise LD on LG11
Map length: 111.5 cM
r2
0
0.2
0.4
0.6
0.8
1

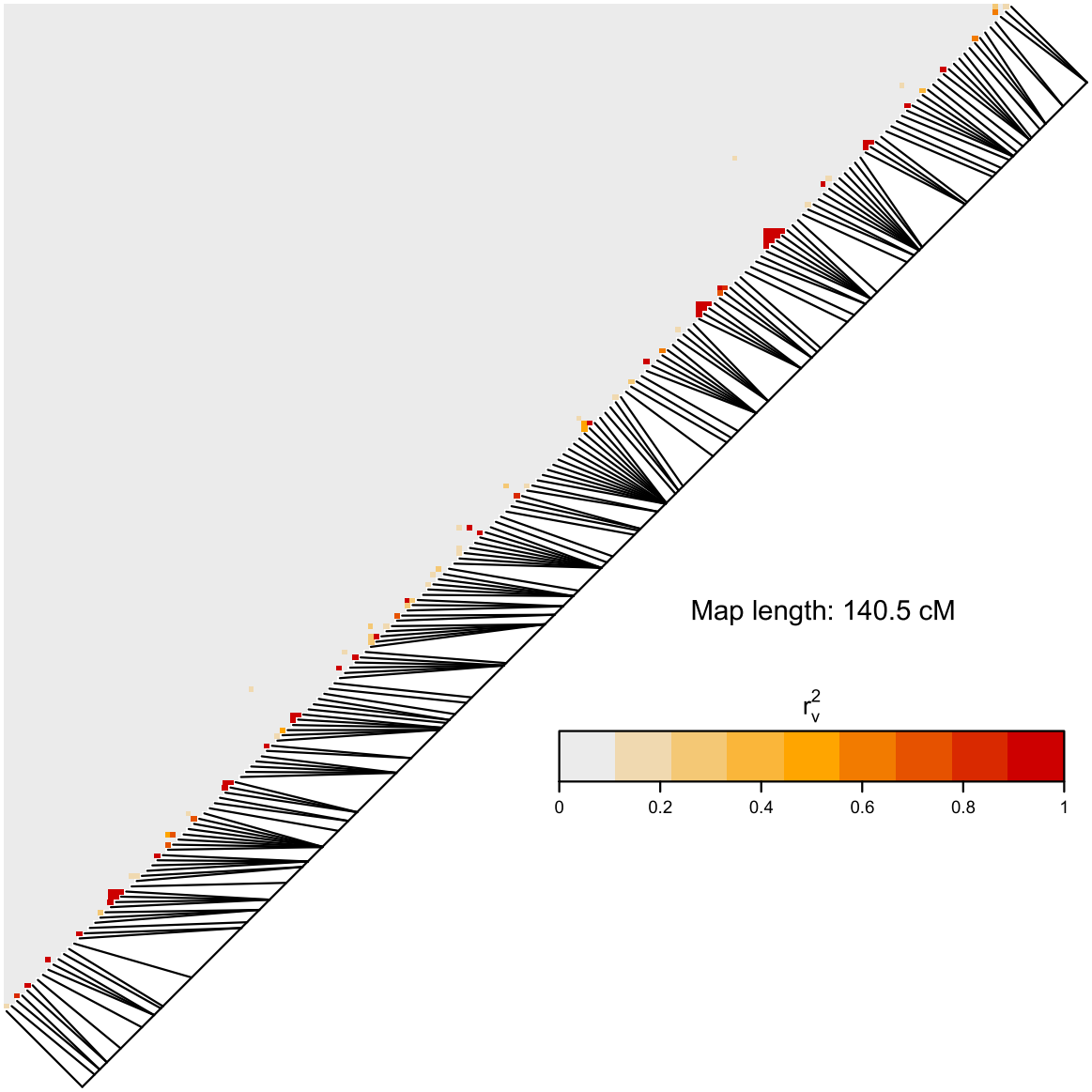

Pairwise LD on LG12
Map length: 144.4 cM
r2
0
0.2
0.4
0.6
0.8
1

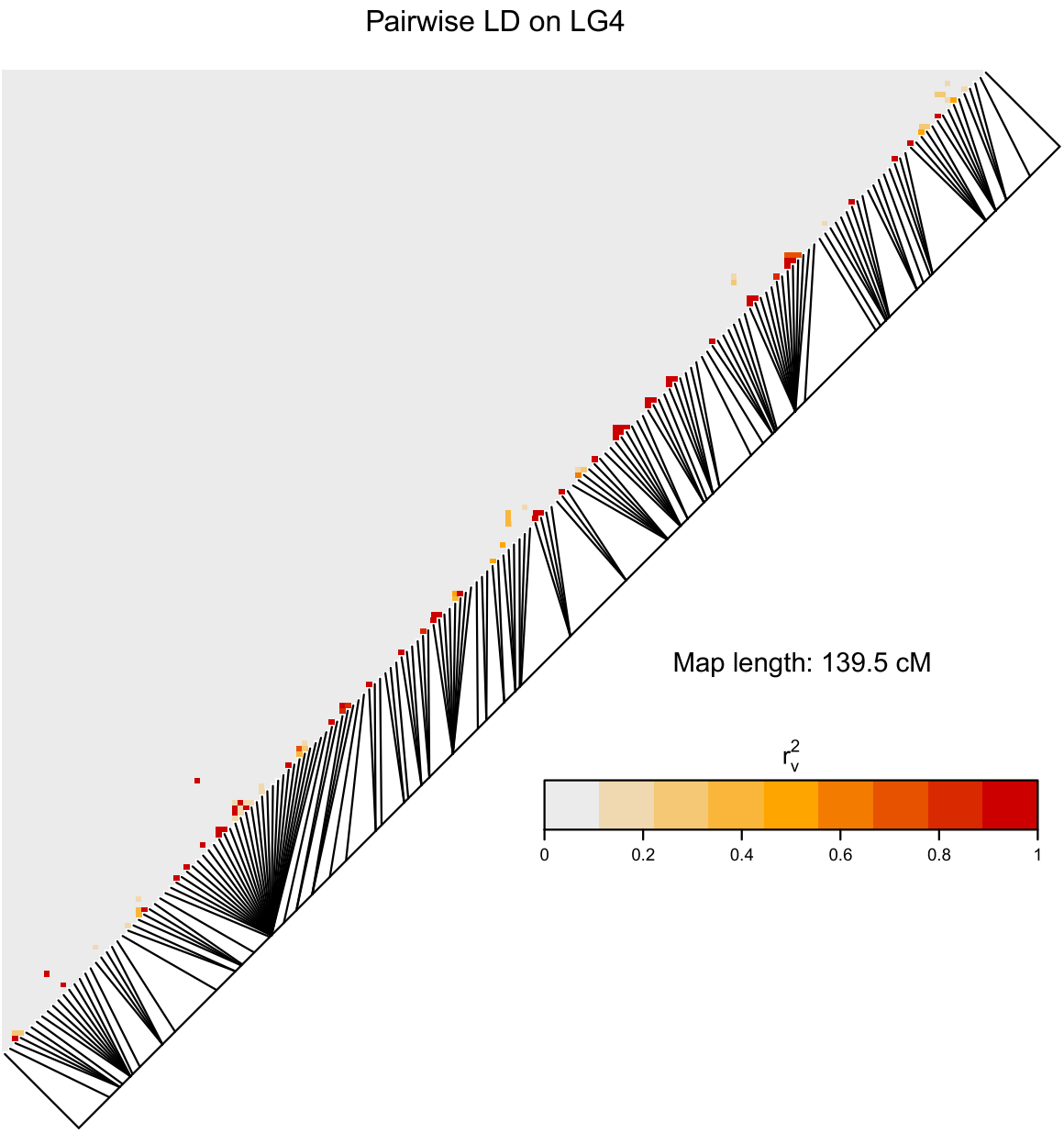
supplemental S4B

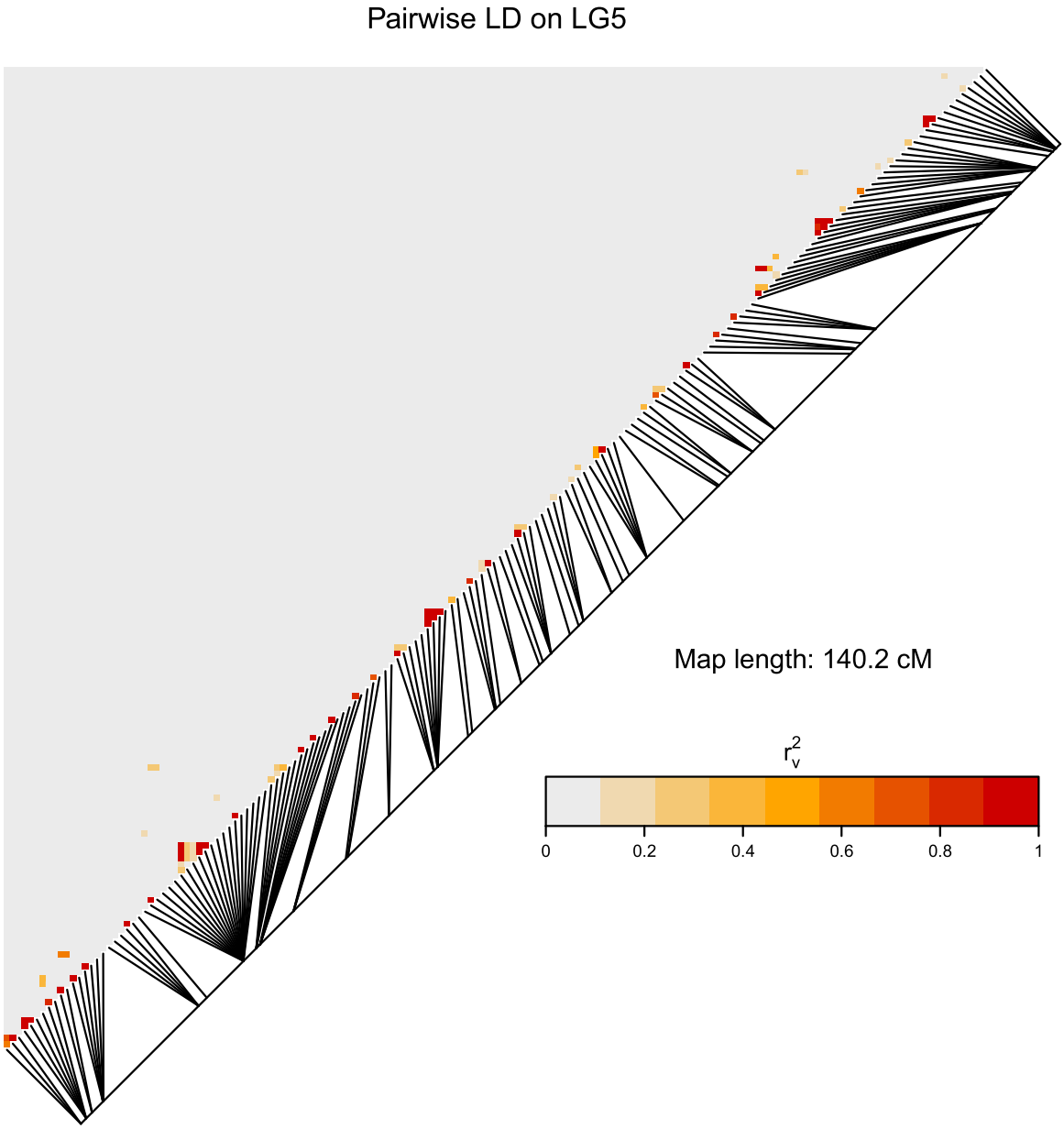



Pairwise LD on LG2
Map length: 153.2 cM
r^2_v
0
0.2
0.4
0.6
0.8
1

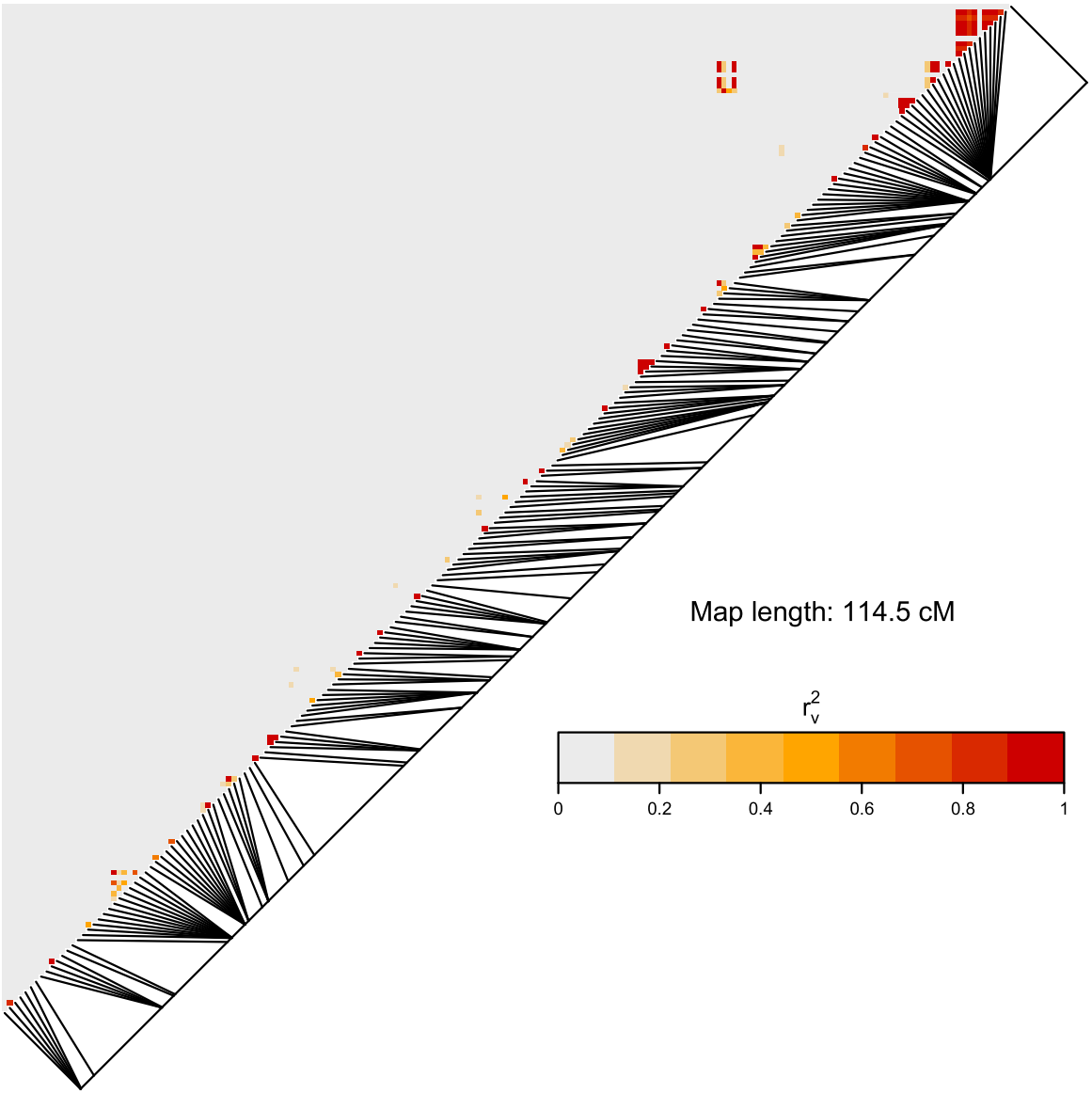
Pairwise LD on LG3

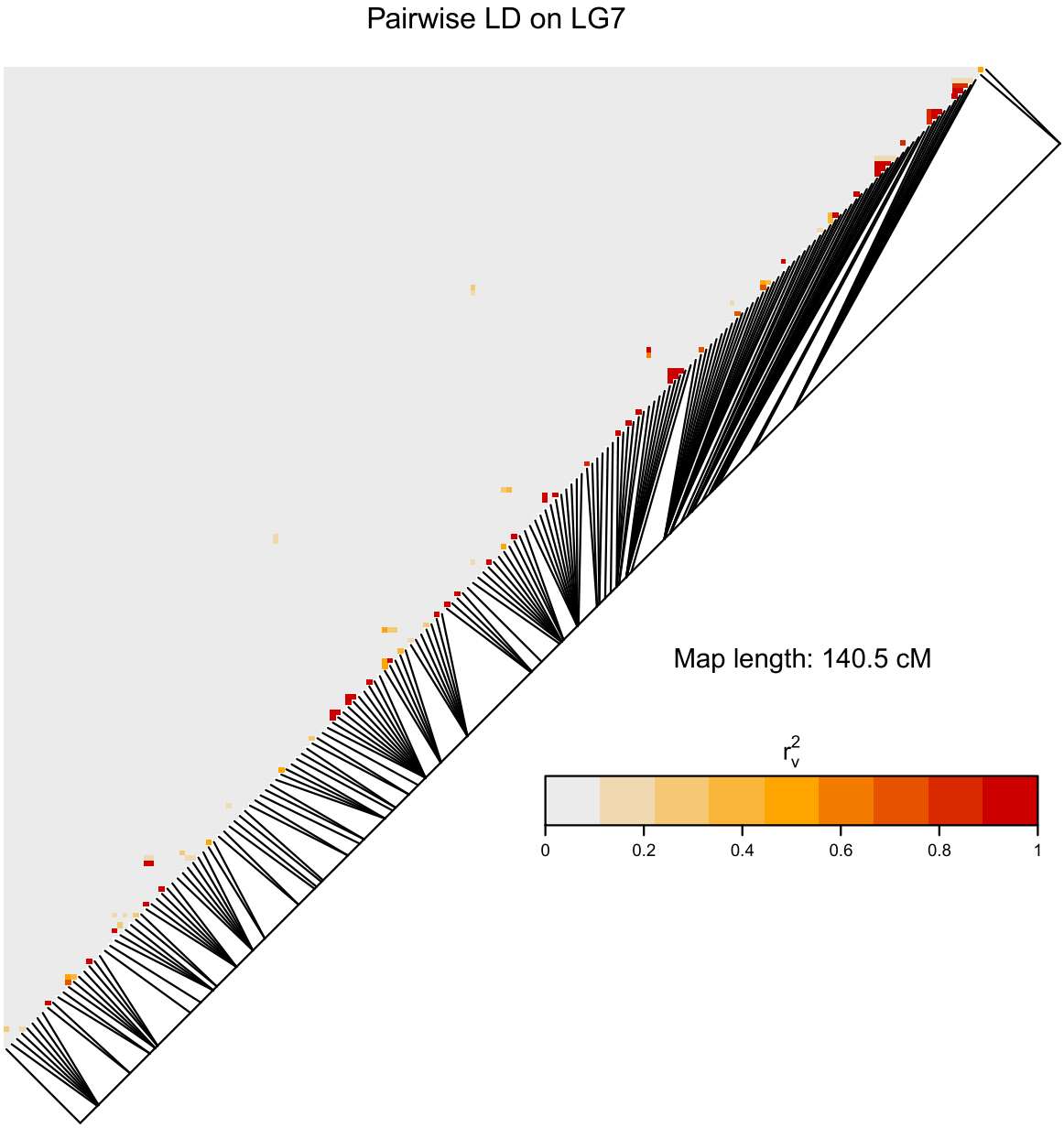


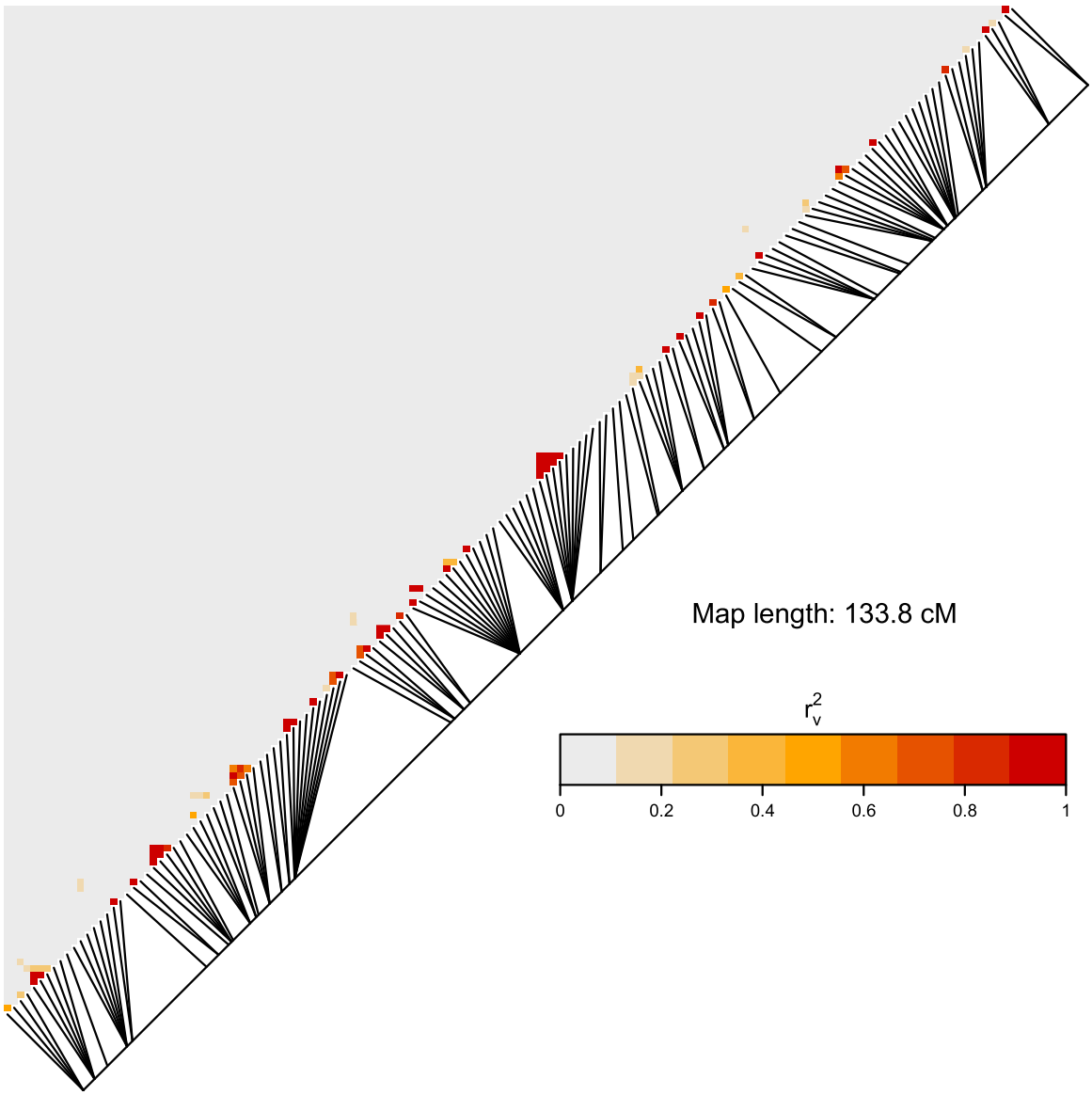

Pairwise LD on LG4
Map length: 139.5 cM
r^2_v
0
0.2
0.4
0.6
0.8
1


Pairwise LD on LG5
Map length: 140.2 cM
r^2_v
0
0.2
0.4
0.6
0.8
1

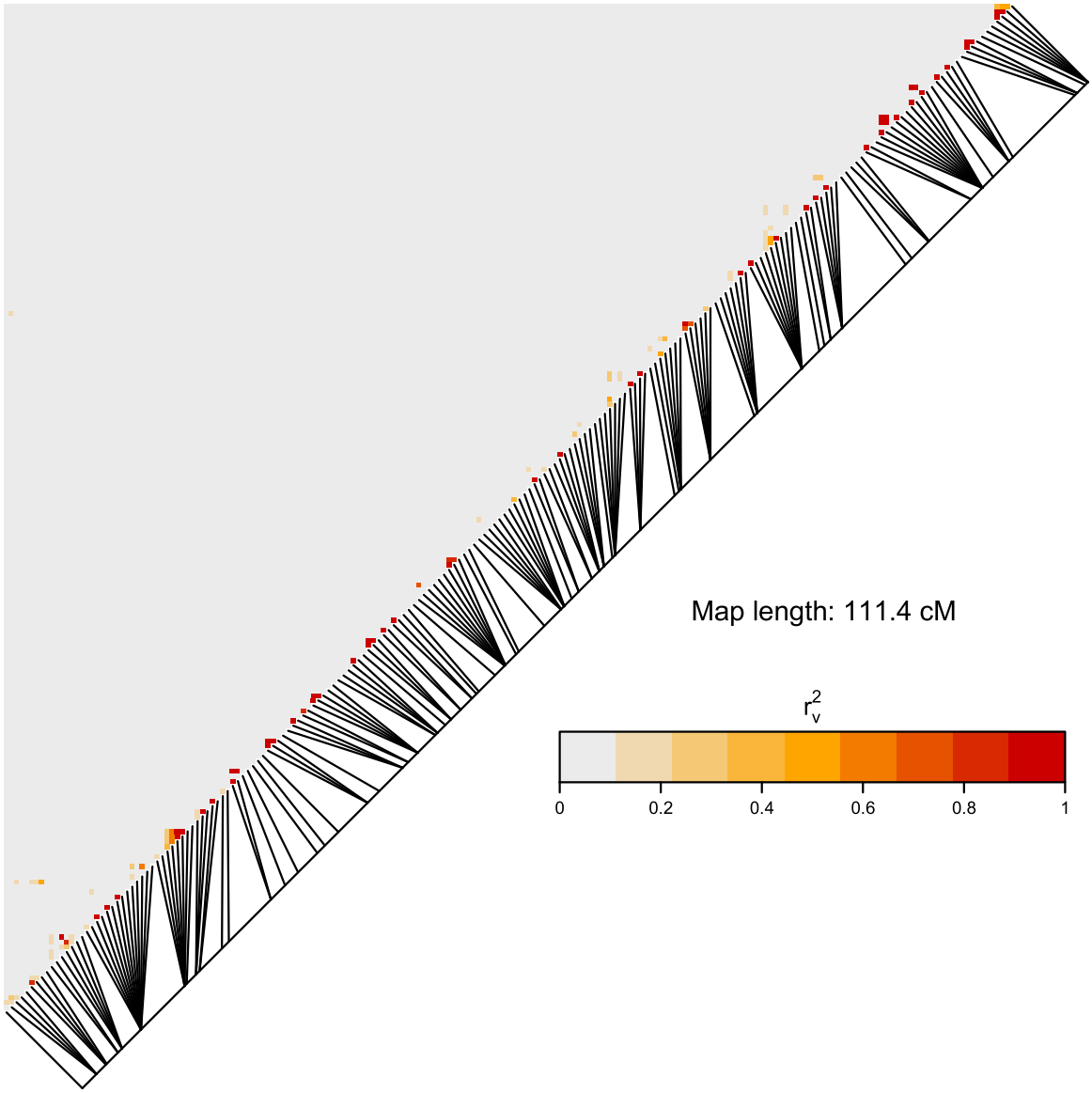
Pairwise LD on LG6



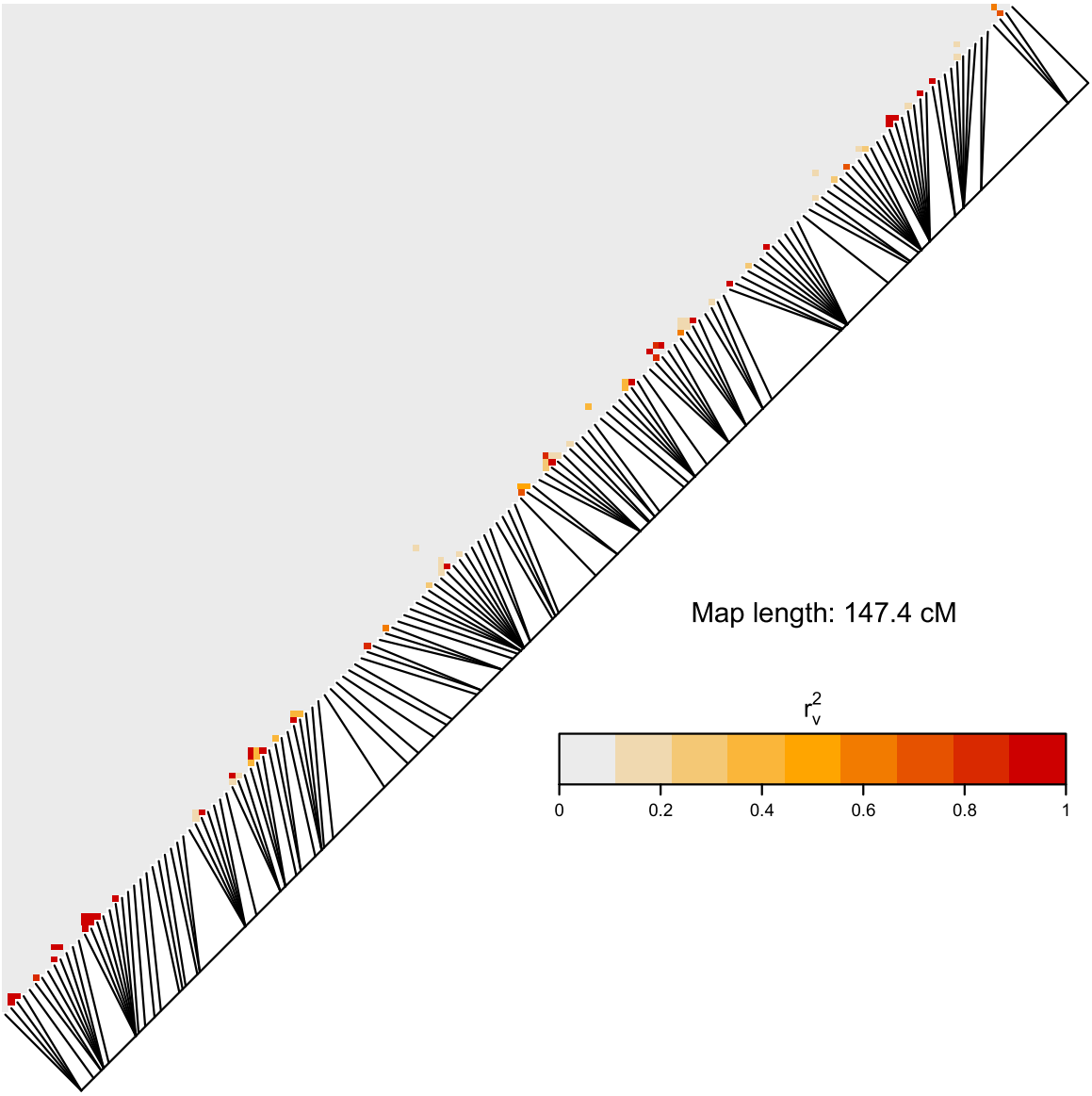

Pairwise LD on LG7
Map length: 140.5 cM
r^2_v
0
0.2
0.4
0.6
0.8
1


Pairwise LD on LG8
Map length: 133.8 cM
r^2_v
0
0.2
0.4
0.6
0.8
1

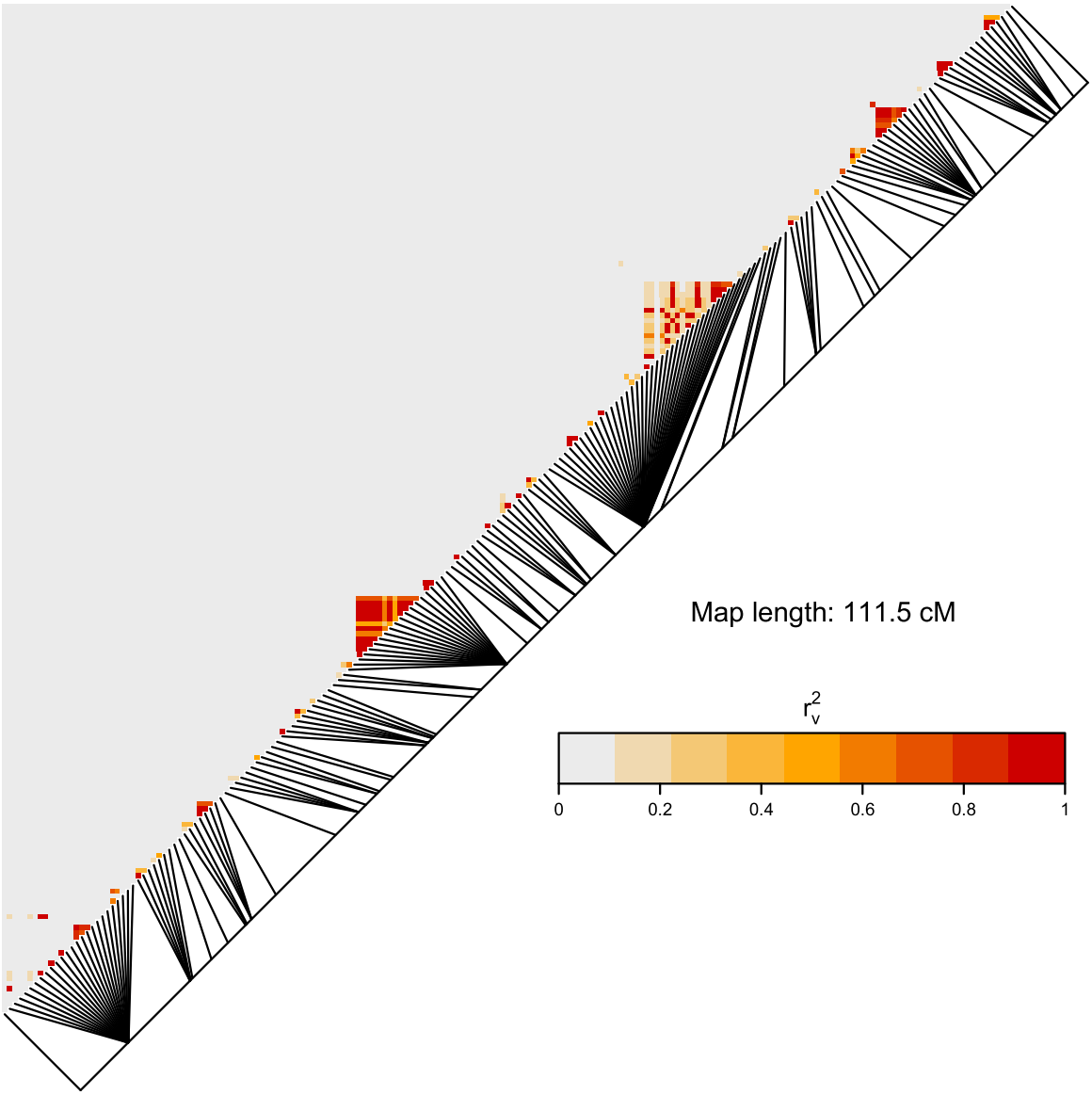
Pairwise LD on LG9



Pairwise LD on LG10



Pairwise LD on LG11



Pairwise LD on LG12