Supplementary Material: Introgressed *Manihot glaziovii* Alleles in Modern Cassava Germplasm Benefit Important Traits and Are Under Balancing Selection

Marnin D. Wolfe, Guillaume J. Bauchet, Ariel W. Chan, Roberto Lozano, Punna Ramu, Chiedozie Egesi, Robert Kawuki, Peter Kulakow, Ismail Rabbi and Jean-Luc Jannink

Supplementary Tables

Supplementary Tables are included as worksheets in the file **SupplementaryTables.xlsx**.

Table S1: Introgression Diagnostic Markers. For each marker in the dataset, indicates chromosome, position, genetic map coordinates (cM) and whether SNP is an IDM (Strict, GlazPoly or tagIDM) or not (nonIDM). Also records *M. glaziovii* allele (GlazAllele) at IDMs. The frequency of the major allele in the *M. esculenta* (EscFreq) and *M. glaziovii* (GlazFreq) reference panels was used to identify loci with absolute allele frequency difference of was used to identify IDM.

Table S2: Raw per individual introgression summary. Summary statistics for each cassava clone are given in a long form with the per individual metric indicated (Metric) along with its value (Value). Per individual metrics: M. glaziovii (PropMg) and deleterious (PropDeleterious) allele frequencies, *M. glaziovii* (PropHomMg) and deleterious (PropHomDeleterious) homozygote frequencies. Regions: per individual summaries were calculated either across all SNPs (GenomeWide) or only for SNPs in the introgression regions on chrs 1 (>25Mb) and 4 (5-25Mb). The cumulative number of field plots (NfieldPlots) planted per clone is given where known. Individual accessions (IID) and summaries are given on a per group (Group) basis.

Table S3: Population level summary. Summary statistics for each cassava population. Per population (Group) statistics (Stat: min, mean, max) for each of the following per individual metrics: *M. glaziovii* (PropMg) and deleterious (PropDeleterious) allele frequencies, *M. glaziovii* (PropHomMg) and deleterious (PropHomDeleterious) homozygote frequencies. Regions: per population summary of per individual metrics calculated either across all SNPs (GenomeWide) or only for SNPs in the introgression regions on chrs 1 (>25Mb) and 4 (5-25Mb).

Table S4: Correlation of Kinships. The correlation between the lower triangular off-diagonals of pairs of kinship matrices (KinshipCoefficient) calculated with different combinations of markers according to the variables Matrix1 and Matrix2. Different criteria (Criterion: LDscore, maxLD or None) with varying thresholds (Threshold) of the corresponding criterion for inclusion of SNPs in the kinship matrices. Matrices correlated were constructed either from IDM markers only, IDM+tagIDM markers or nonIDM markers only.

Table S5: Field Trial Summary Summary for each trait in each field trial that we analyzed. In addition to institute, location, year harvested, trial name and trial type, meta information about the numbers of observations (Nobs), clones (Nclone) and replications (Nrep) are provided as well as the Nobs-to-Nclone ratio (ObsToCloneRatio).

Table S6: Per-trial Model Summary For each trait-trial combination, we fit several models, as described in the text. The p-values from likelihood ratio tests for significant IDM (LRT_IDM) and non-IDM (LRT_nonIDM) are given. AIC-based model comparisons were made and the best (BestModel) and next-to-best (SecondBestModel) models are indicated. The AIC difference between the best and second-best models (AICdiffFrom2ndBest) is shown. The correlation-of-estimates from the two component PARTITIONED model (CorOfEsts_IDMvsNonIDM) is also included. Finally, logical variables indicating whether or not a trial passed three preliminary filters described in the text are also provided (Filter1, Filter2, Filter3).

Table S7: Random partitions. We used three random partitions of the set of SNPs in our dataset as a point of comparison to the IDM vs. non-IDM partition which is the focus of our study. This table gives, for each SNP, the status as either IDM or nonIDM, which was assigned at random in each of three samples. These should be distinguished from actual status as IDM, which is indicated elsewhere based upon allele frequencies as described in the text.

Table S8: Cor. of randomly partitioned GRMs The correlation of the lower triangular off-diagonals of kinship matrices (CorGRMs) made with IDM-sized sets of SNPs to kinships made with the non-IDM sized set. The sets of SNPs in each pair of kinship (each row) was partitioned either at random or based on true IDM-status (SampleMethod).

Table S9: LDAK Weights. Weights and related raw output from the software LDAK.

Table S10: Per-trial Per-model LogLik, AIC, VarComps, Etc. For each trait-trial combination, we fit several models (Models), as described in the text. This table gives detailed output from each model fit for each trait-trial analysis. Included are the log-likelihood (LL), number of parameters (Npar), akaike information criterion (AIC), the variance component (Var) from each relevant source of variance (Source), the percent variance explained by the variance component (PVE) and a likelihood ratio test against the null model

(no genetic variance; LRT_null).

Table S11: Summary of curated multi-trial data After combining all trait-trial data passing filters described in the test, we combined data within institute for each trait for a join multi-environment analysis. This table indicates the sample size (Nobs) along with numbers of clones (Nclone), trials (Ntrial), replications (Ntrial_rep) and a summary of the distribution of observations (1st, and 3rd quartiles, min, max, mean and median) per clone (nPerGeno suffix).

Table S12: Summary of multi-trial analyses Gives the p-value for the likelihood ratio test for the significance of the IDM vs. non-IDM partition of genetic variance (LRT_partition). Whether the ALL or PARTITIONED model is better or the same is given based on which had a smaller AIC and whether there were more than 2 units AIC difference respectively (PARTvsALL). AIC comparison for the PARTITIONED vs. the IDMnull models are also given (PARTvsIDMnull). Heritability of the IDM and nonIDM parts (h2_IDM and h2_nonIDM) are also shown. Results are given for both IDM-based and random partitions, with and without LDAK LD weighting of the kinship matrices.

Table S13: Multi-trial Per-model LogLik, AIC, VarComps, Etc. For each trait, results from each model fit to each multi-trial dataset are given. Kinship matrices (GRM) used in the estimation of genetic variances were either LDAK LD weighted or not. SNPs used in constructing the GRMs were partitioned either based on true IDM-status or at random (SampleMethod). The p-value from the likelihood ratio test of the partitioned vs. the not-partitioned models (LRT_partition) and whether or not that LRT was significant (p-value < 0.05) are shown (LRTpartitionSig). AIC-based comparisons of partitioned to the not-partitioned (PARTvsALL), and partitioned to the model without the IDM component (PARTvsIDMnull) are summarized. Model comparisons were designated as either the SAME (AIC difference <=2) or as one or the other model being better (AIC difference >2). Finally, the heritability (percent variance explained) by the IDM and non-IDM variance components are given (h2_IDM, h2_nonIDM).

Table S14: Compare IDM partitioned to randomly partitioned multi-trial models. For each trait and each of three random partitions of the genome, compare to the IDM-based genome partition based on the akaike information criterion (AICdiff = AIC_idms - AIC_random). IDMvsRANDOM summarizes whether the models were essentially the same (SAME, -2<AICdiff<+2) or either the IDM or the random-partitioned models (IDMbest if AICdiff < -2, RANDbest if AICdiff > 2). The p-value from the likelihood ratio test on the significance of the variance component associated with the IDM-sized set of SNPs is given for each model (pIDM, pRandom) and signficance was declared if the p-value < 0.05 (IDMsig, RandomSig).

Table S15: Correlation of kinships constructed with vs. without LD-adjustment. Correlate the lower-triangular off-diagonals (CorOffDiag) and the diagonals (CorDiag) of the LDAK weighted and unweighted kinship matrices. Compare matrices constructed by rrBLUP packages A.mat function (unweighted) to the LD-adjusted (weighted) matrix constructed by the software LDAK with either ALL, IDM, or nonIDM SNPs (Category).

Table S16: Correlation of IDM and non-IDM kinships constructed either with or else without LD-adjustment. Correlate the lower-triangular off-diagonals (corOffDiag) and the diagonals (corDiag) of the IDM vs. nonIDM (IDMvREST) kinship matrices. Compare matrices constructed by rrBLUP packages A.mat function (unweighted) to the LD-adjusted (weighted) matrix constructed by the software LDAK.

Table S17: Compare multi-trial analyses with vs. without LD-adjustment. For each Trait-Institute multi-trial dataset, compare the results from models using the LDAK weighted kinship matrix to those using an unweighted kinship matrix. Comparison was made primarily on the basis of change in the heritability (percent variance explained) of the IDM component (ChangeIDM = h2_weighted - h2_unweighted). The AIC for weighted vs. unweighted analyses are given and the AICdiff is expressed as AIC_ldweighted - AIC_unweighted). The model with the lowest AIC by at least 2 points was declared the best (BestGRM). Whether or not (TRUE vs. FALSE) the LRT on the IDM variance component was significant (p-value < 0.05) for the weighted and unweighted models is also shown.

Table S18: Summary of GWAS results For each Trait and Type of GWAS, we summarize bonferroni-significant results on a per chromosome basis. Indicates the number (Nsig), min and max position (minPos, maxPos) and the *M. glaziovii*-allele effect (Mean, Min, Max).
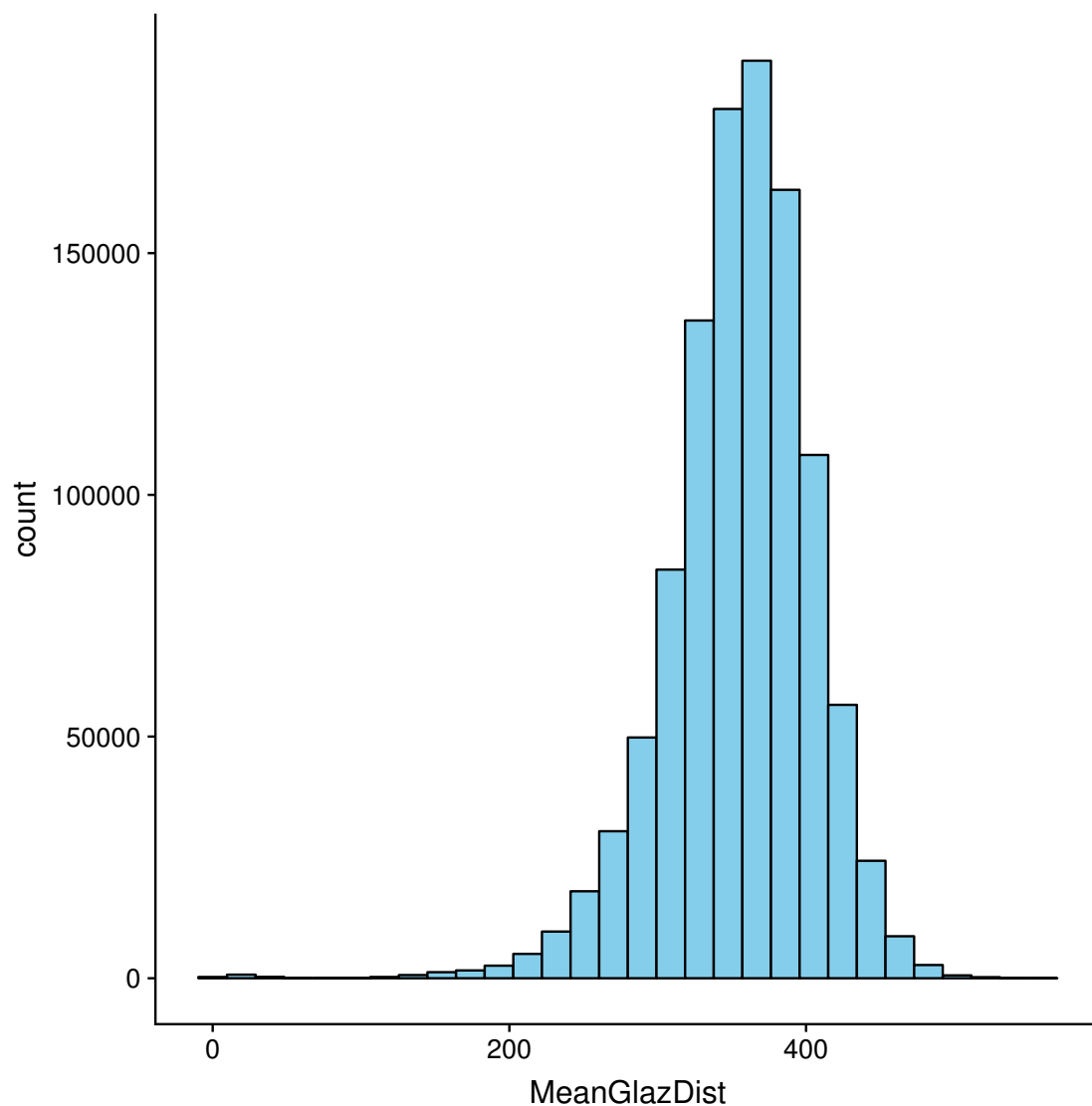
Table S19: Cross-validation results Prediction accuracy from each rep of five-fold cross-validation conducted on each Trait-Institute-SampleMethod-Model and calculated for each variance component where relevant. Accuracy was measured as the pearson correlation between test-set GEBV and validation BLUPs. SampleMethod indicates whether the genome was partitioned based on true IDM-status of SNPs or else whether assignment was made at random. Models fit either covered the entire genome, unpartitioned (ALL), nonIDM only, IDM only or both at the same time (PARTITIONED).

Table S20: Summary of cross-val. results Cross-validation accuracy results summarized for each Trait-Institute-SampleMethod. The mean (across reps) differences in accuracy between models is given (DiffAcc_PartitionVSnot, DiffAcc_PartitionMinusNonIDMonly, DiffAcc_AllminusWithoutIDMsnp).
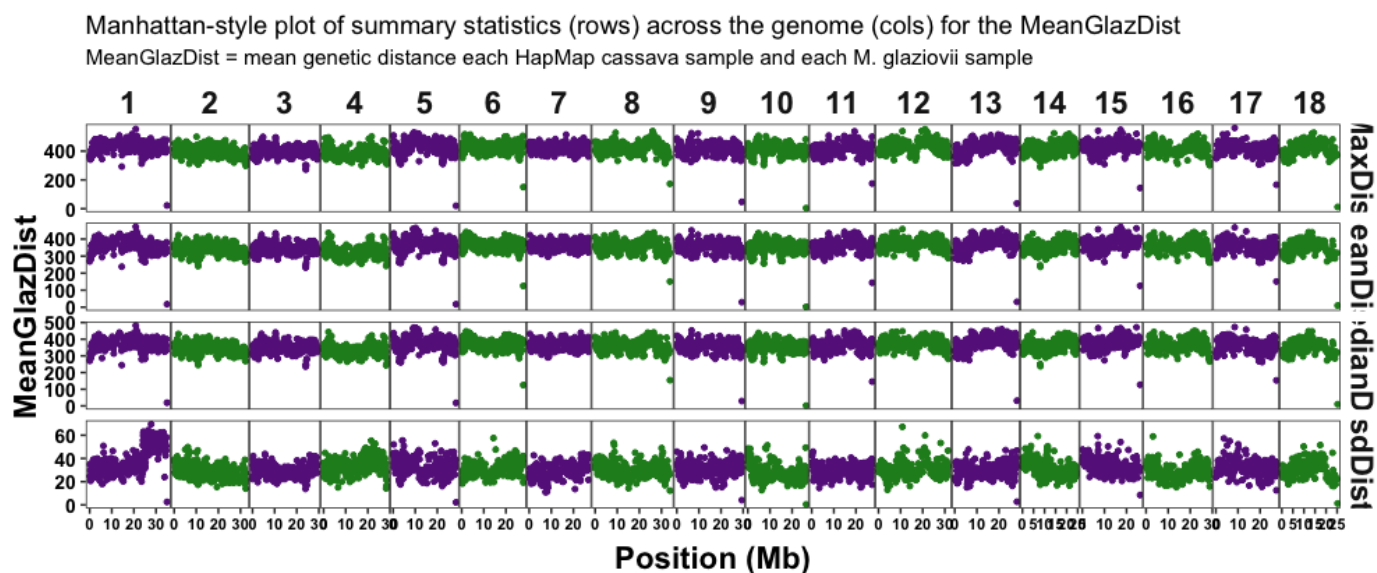
Table S21: Summary of genetic distance per megabase. The cumulative genetic distance in centimorgans for each one megabase region along the genome, highling major introgression regions (Chr1_IntrogRegion: distal 25Mb of chr. 1, Chr4_IntrogRegion: from
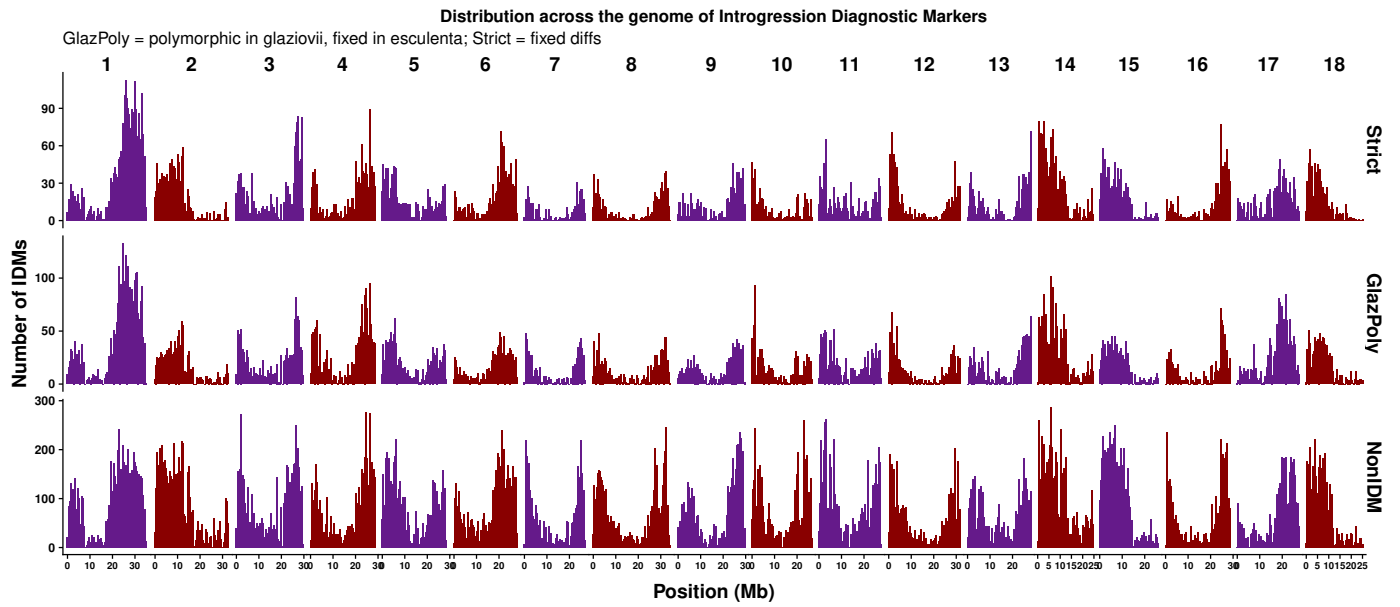
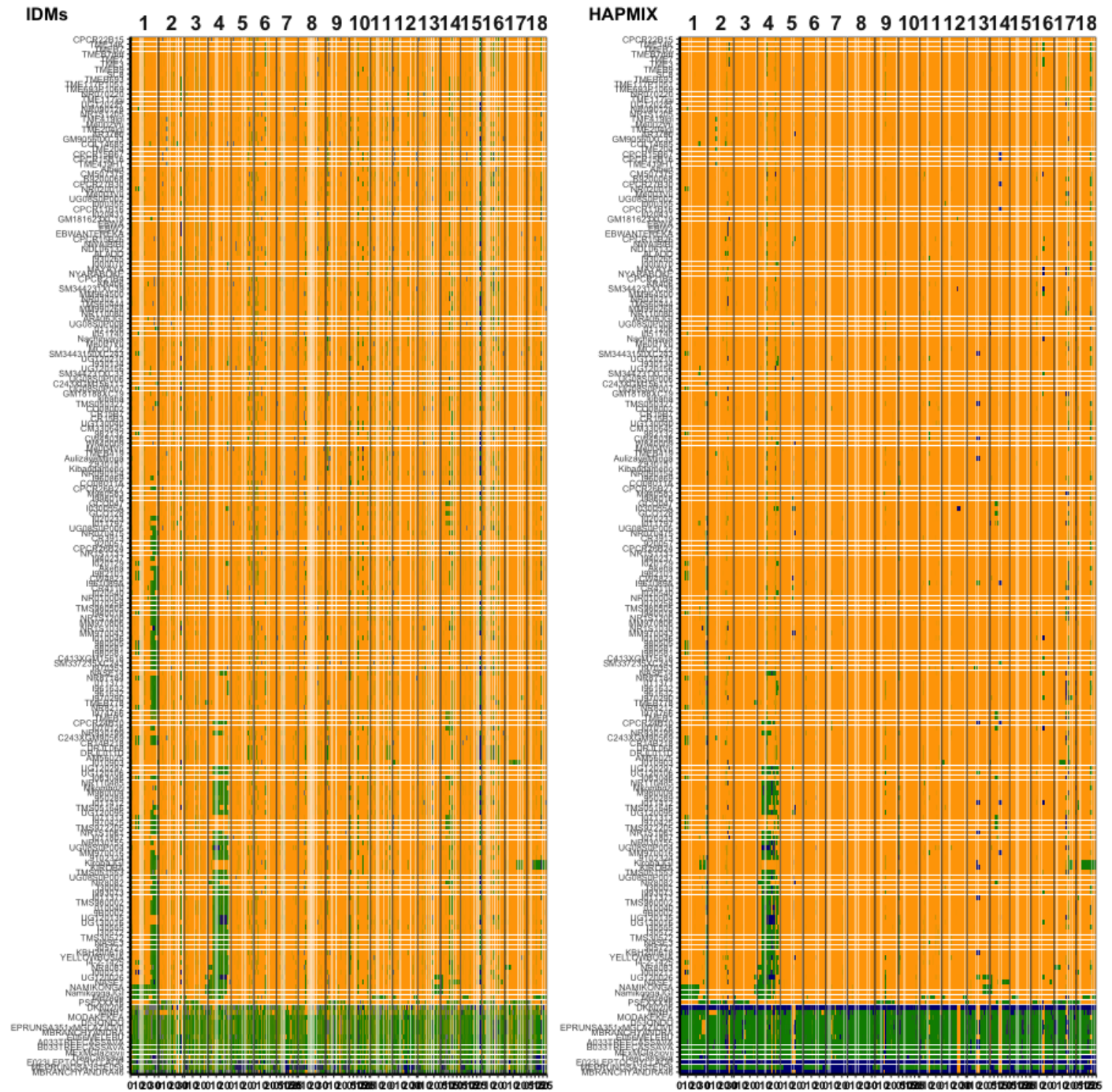5-25Mb on chr. 4, RestOfGenome).

Supplementary Figures

Figure S1: Distribution of MeanGlazDist in the cassava portion of HapMapII. MeanGlazDist = mean Hamming distance for a 1000 SNP window between the *M. glaziovii* reference panel and each cultivated-cassava sample.

Figure S2: Manhattan-style plot of summary statistics (rows) across the genome (cols) for the MeanGlazDist. MeanGlazDist = mean Hamming distance for a 1000 SNP window between the *M. glaziovii* reference panel and each cultivated-cassava sample.
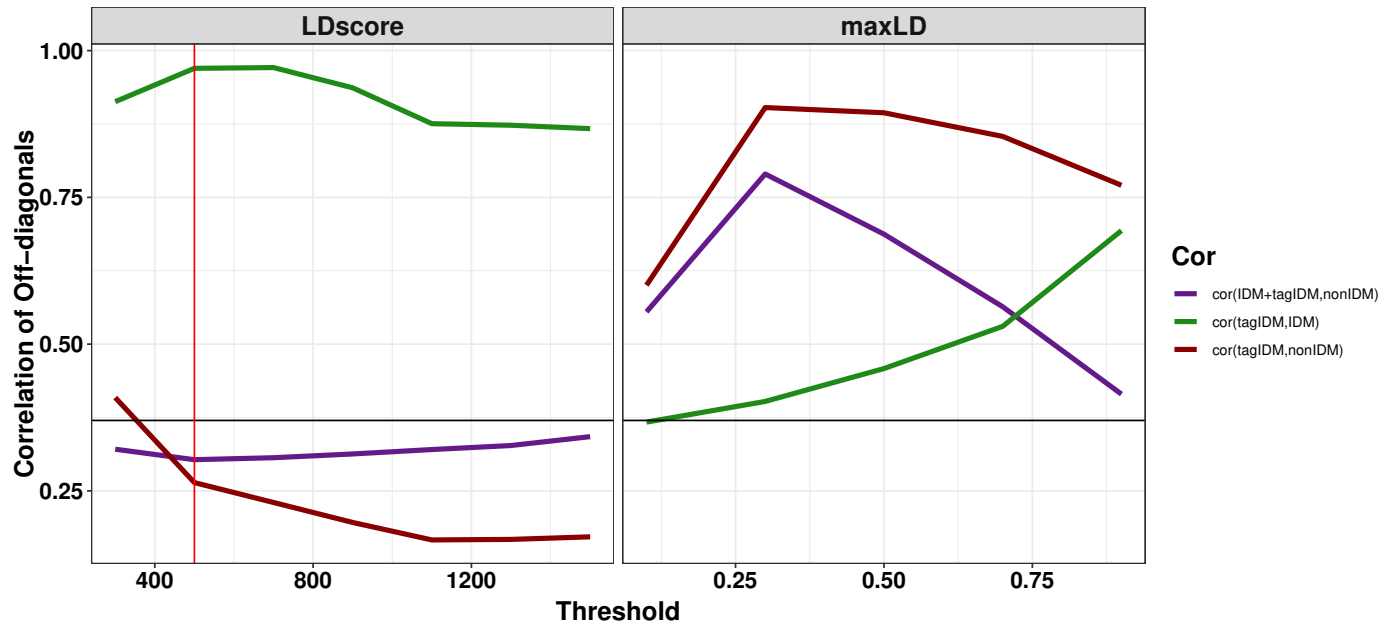
Figure S3: Genome-wide distribution of introgression diagnostic markers (IDMs) relative to the rest. We note from the plot above that the distribution and density of both *Strict* and *GlazPoly*, are very similar to each other *and* also somewhat similar to the *nonIDM* SNPs. The important thing here is that we see that the genome coverage of the diagnostic markers is similar to that of GBS markers in general. We will only be able to reliably call admixed cassava germplasm as introgressed in regions where we have *IDMs*.

Figure S4: Comparison of introgressions detected by IDMs vs. HAPMIX in HapMapII. The mean M. glaziovii allele dosage at IDMs in 250Kb windows across the genome is based on introgression diagnostic markers (IDMs, left) and also based on local admixture modelling (HAPMIX, right). Physical position on each chromosome is depicted in megabases (Mb) along the x-axis. Colors range from orange (0 M.g. alleles), to green (1 M.g. allele), to dark blue (2 M.g. alleles). Rows (clones) are aligned across A and B and sorted within based on the IDM-based genome-wide proportion *M. glaziovii*.
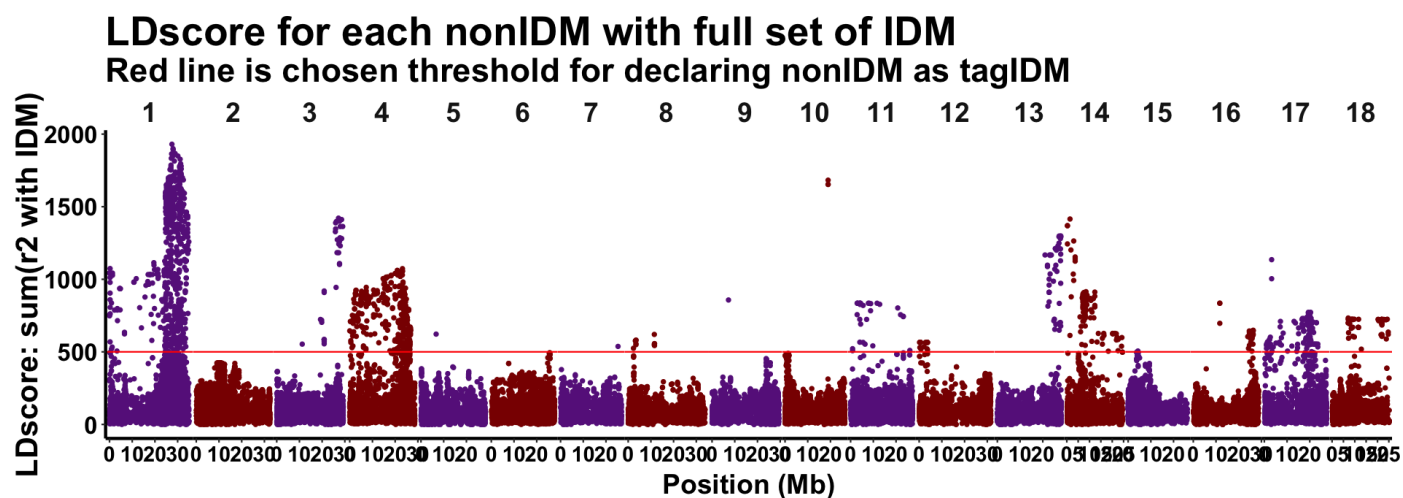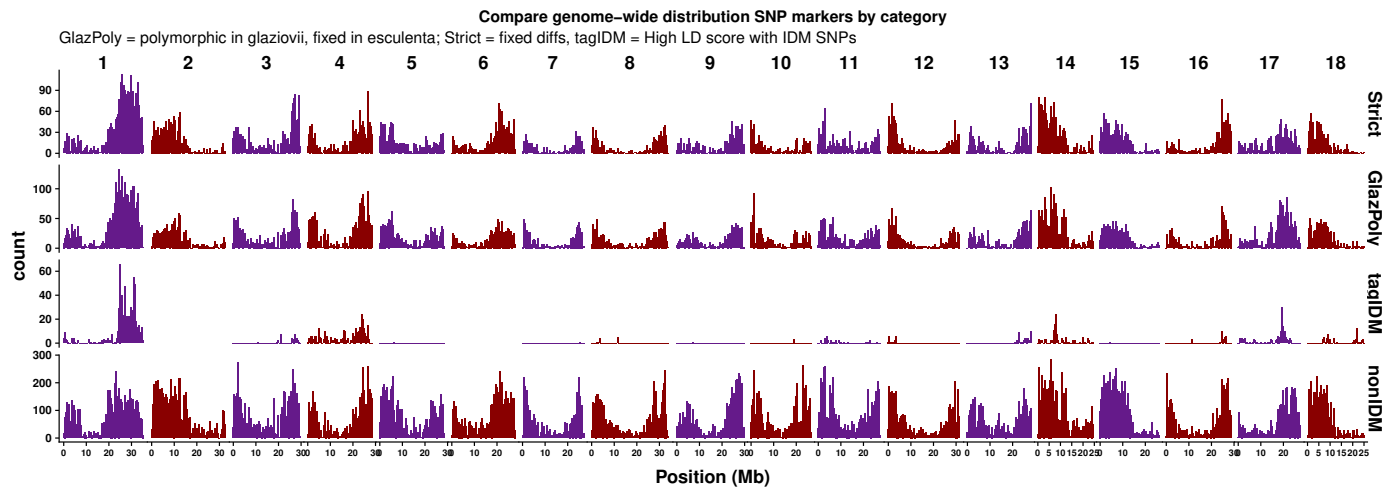
Figure S5: Correlations among the off-diagonals of pairs of kinship matrices. The plot above shows three correlations between the off-diagonals of pairs of kinship matrices created using different sets of SNPs (y-axis). The SNP sets were defined based on a range of thresholds (x-axis) for two criteria for thresholding (panels). The correlations are: cor(IDM+tagIDM,nonIDM), which we want to minimize (purple), cor(tagIDM,IDM) which we want to maximize (green) and cor(tagIDM,nonIDM), which we also want to minimize (red). The horizontal line is the correlation (0.37) between the kinships created for the two original categories (IDM vs. nonIDM), without considering LD. The vertical red line indicates the threshold and the criterion that we ultimately chose to use in order to define addition nonIDM SNPs that "tagged" the introgression regions. The maxLD criterion seems to generally produce correlations between IDM and nonIDM that are greater than the original categories, therefore we rejected that option. In contrast, the LDscore criterion did very well. The cor(IDM+tagIDM,nonIDM) was always lower than the unmodified categories, the cor(tagIDM,IDM) increased to nearly 1 with LDscore thresholds <600 but still remained high across the range and the cor(tagIDM,nonIDM) dropped very low as LDscore-threshold was increased. We chose to use the LDscore criteria with a threshold such that nonIDM SNPs with LDscore>500 were designated as tagIDM and included with the IDM SNPs in downstream analysis (calculation of kinship matrices for variance partitioning).
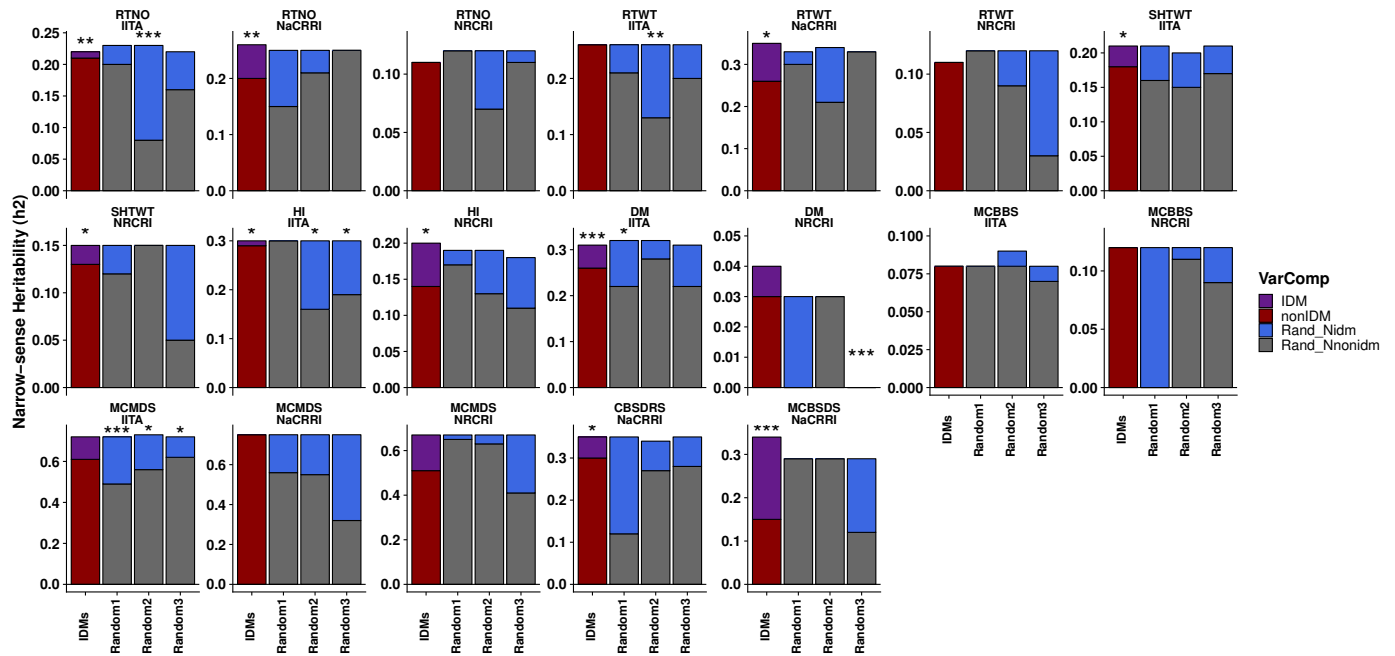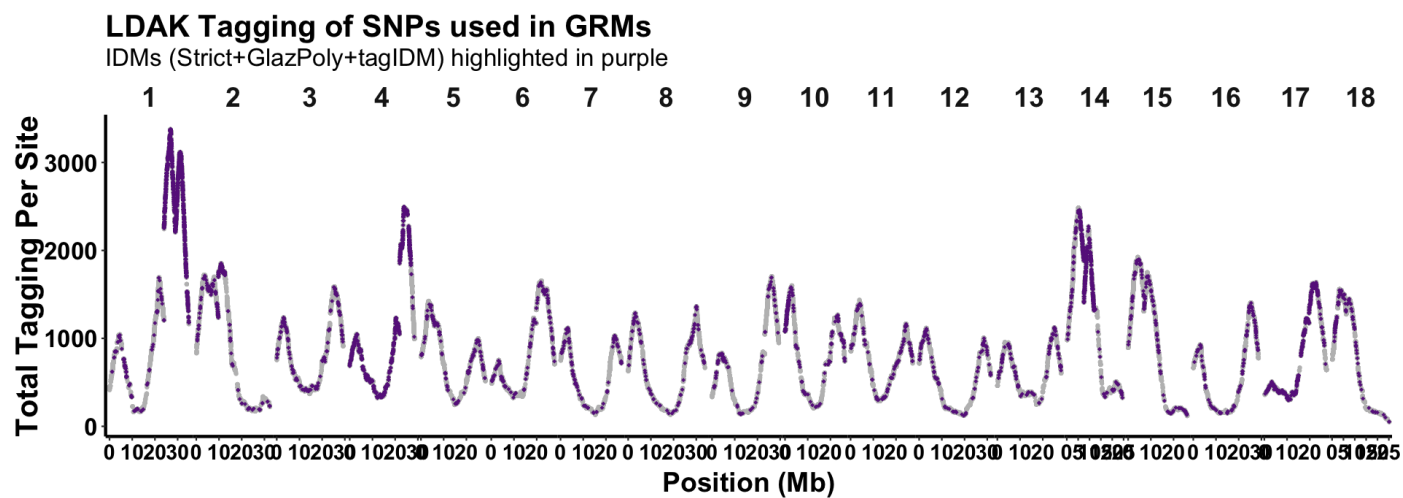
Figure S6: LDscore for each non-IDM with the full set of IDM. The plot above shows the LDscore's (sum of r^{2}_{LD}) of nonIDM SNPs (y-axis) versus their position in the genome (x-axis). The red horizontal line indicates the threshold above which nonIDM SNPs were designated tagIDM.

Figure S7: Genome-wide distribution of introgression diagnostic markers (IDMs) including tag-IDMs. Shows the distribution and density along the genome of IDMs both *Strict* and *GlazPoly* as well as *tag-IDMs* and *non-IDM* SNPs.
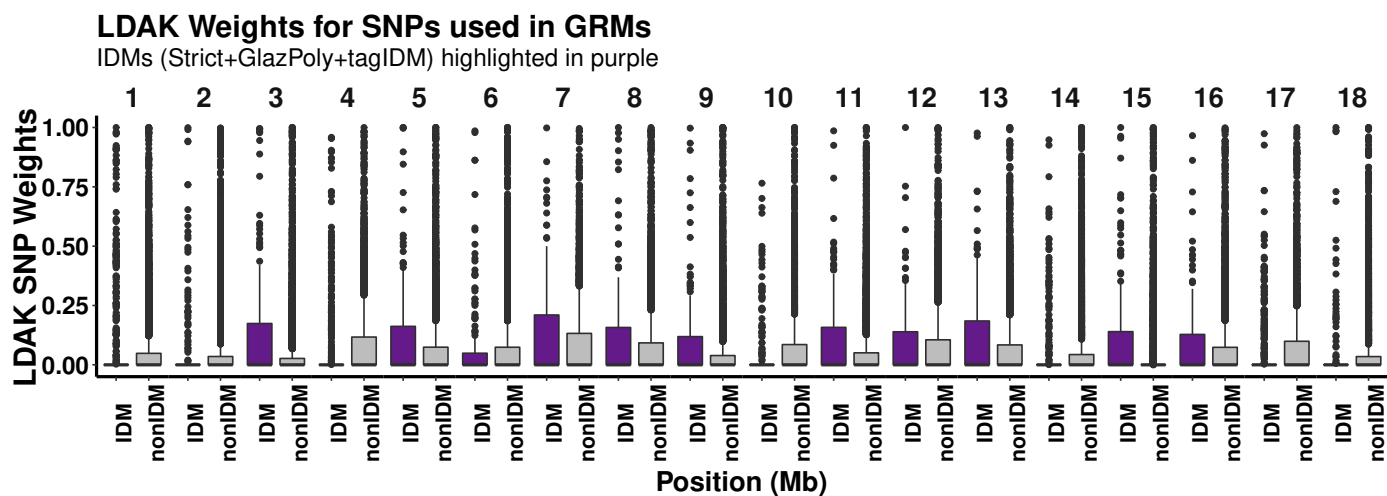
Figure S8: Heritability attributable to IDM vs Random genome-partitions. Each panel shows, for one trait measured at one breeding program, the heritability (y-axis) measured partitioning the genome either based on IDMs or else according to 3 random samples of equivalent number (38000) to the IDMs (x-axis). Heritability was estimated from partitioned genomic mixed-models and the portion of heritability attributable to N_{IDM} component (IDM=purple, Random=royal blue) vs. the rest of the genome (IDMs=dark red, Random=gray) is shown. Stars atop bars represent the level of significance in a likelihood ratio test for the significance of the N_{IDM} component (∗ ∗ ∗ p<0.0001, ∗ ∗ p<0.001, ∗ p<0.05).
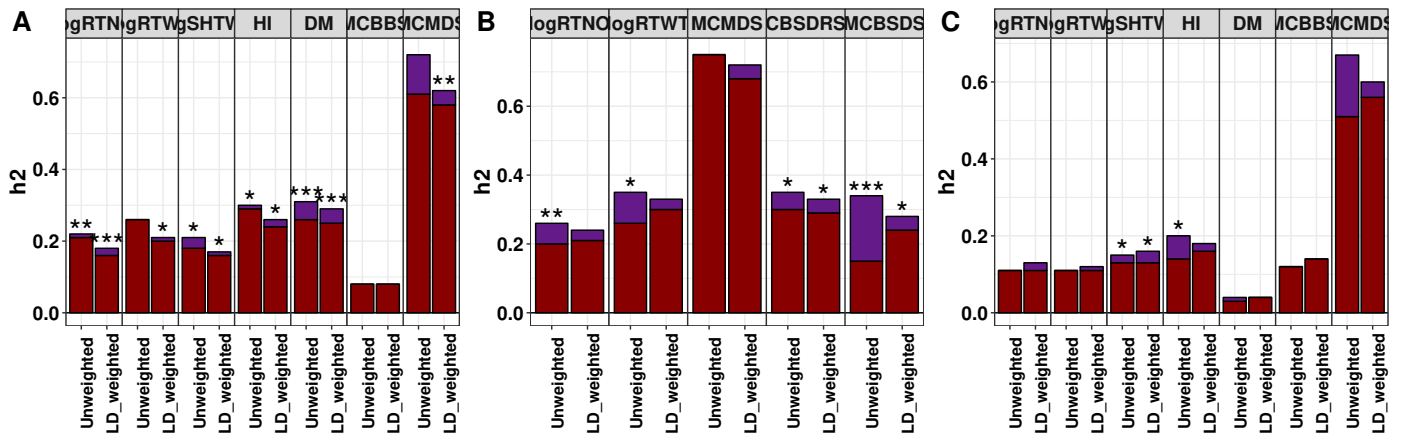
Figure S9: LDAK Tagging of SNPs used in GRMs. Plot showing the LDAK score measuring the total amount of LD for each SNP (y-axis; column 4 from LDAK output) vs. its chromosomal coordinates (megabases, Mb). IDM SNPs (including tag-IDM) are highlighted in purple, non-IDM SNP are gray.

Figure S10: LDAK Weights used for weighting SNPs contribution to GRMs. Plot showing the LDAK weight for each SNP (y-axis) vs. its chromosomal coordinates (megabases, Mb). The weight adjusts the contribution of each SNP to the kinships measured in an LD-adjusted genomic relationship matrix (GRM). IDM SNPs (including tag-IDM) are highlighted in purple, non-IDM SNP are gray.

LDAK Weights for SNPs used in GRMs
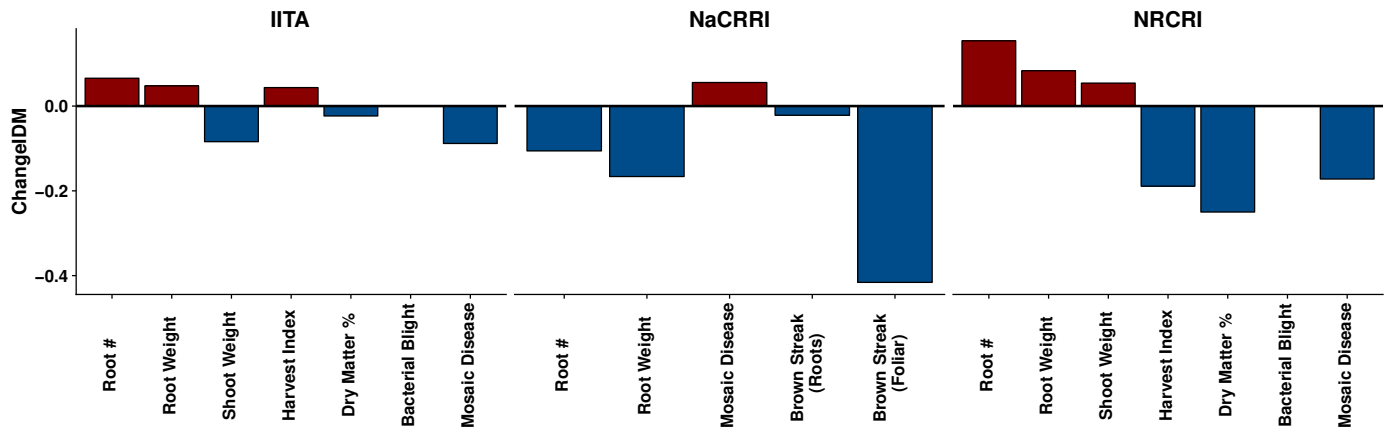IDMs (Strict+GlazPoly+tagIDM) highlighted in purple

Figure S11: LDAK Weights for each chromosome, comparing IDM to non-IDM SNPs. Boxplot shows the LDAK weight for each SNP (y-axis) vs. its status as an IDM SNPs (including tag-IDM) or not (x-axis) for each chromosome (horizontal panels). The LDAK weight adjusts the contribution of each SNP to the kinships measured in an LD-adjusted genomic relationship matrix (GRM).
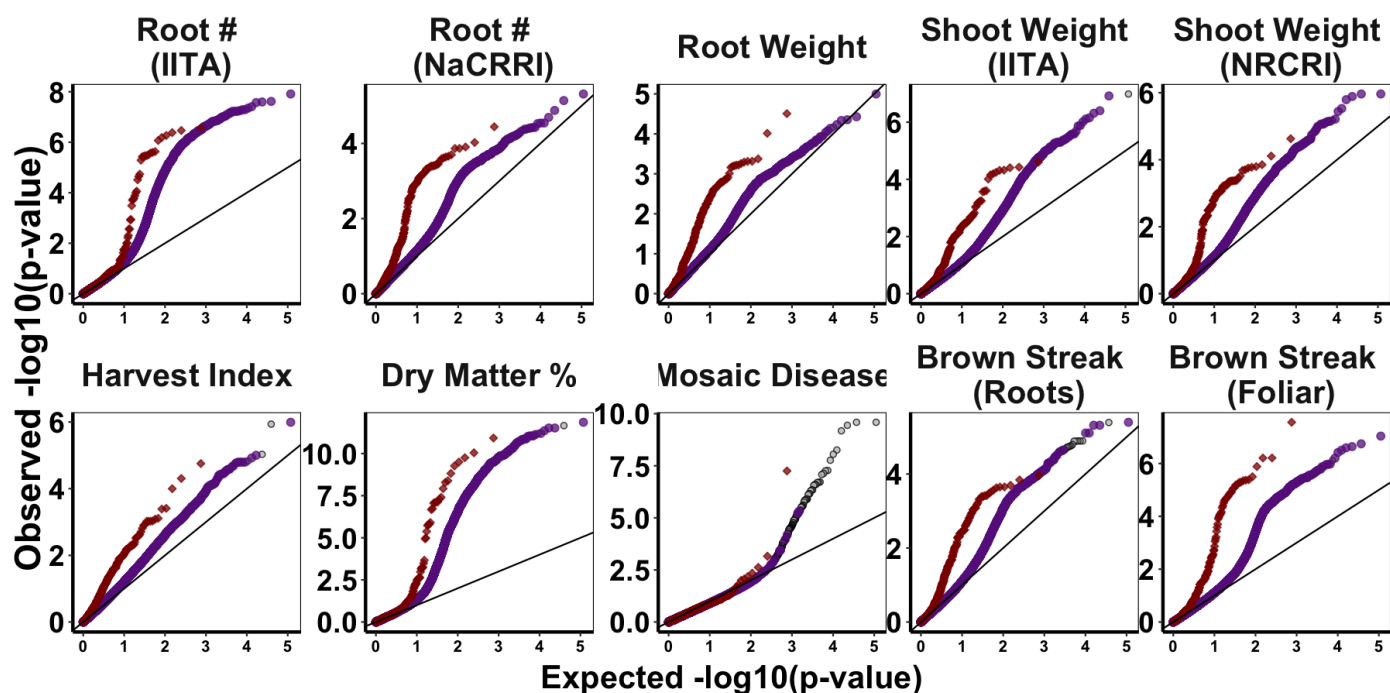
Figure S12: Heritability attributable to introgressions with vs. without LD-adjustment. The heritability (y-axis) of introgression regions for each trait (horizontal panels) is shown for each breeding program (A = IITA. B = NaCRRI. C = NRCRI.). Results for two models are shown (x-axis): one where the genomic relationship matrix (GRM) was LD-adjusted using the LDAK method, and the other without LD-adjustment. In either case, heritability was estimated from partitioned genomic mixed-models and the portion of heritability attributable to introgression regions (purple) vs. the rest of the genome (dark red) is shown. Stars atop bars represent the level of significance in a likelihood ratio test for the significance of the introgression-component ($* * *$ p<0.0001, $* *$ p<0.001, $*$ p<0.05).
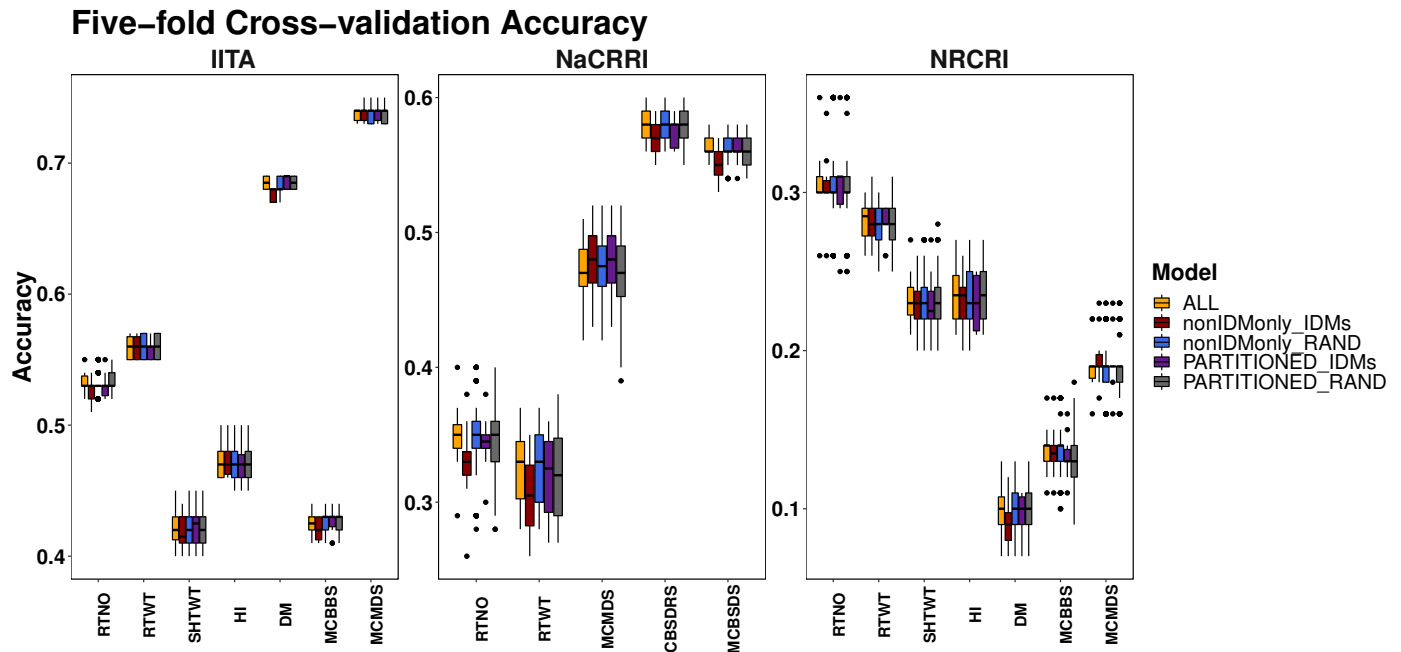
Figure S13: The change in the proportion of the heritability attributable to introgressions due to LD-adjustment of the kinship matrix. For each trait (x-axis) in each breeding program (horizontal panels), the proportion of the total heritability that was due to the introgression component (*Prop_h2IDM*) was computed for two models: one where the genomic relationship matrix (GRM) was LD-adjusted using the LDAK method, and the other without LD-adjustment. The difference between *Prop_h2IDM* between LD-adjusted and un-adjusted estimates (*ChangeIDM*) is plotted on the y-axis. Positive values of *ChangeIDM* (dark red) indicate that *Prop_h2IDM* was larger for the LD-adjusted model, and vice versa for negative values (blue).

Figure S14: Quantile-quantile plots for each introgression-trait associations analysis. Each panel shows QQ-plots for two-types of GWAS for one of 10 Trait-Institute analyses that had bonferroni-significant evidence of introgression-related associations. Two mixed-linear model association analyses are shown, overlayed. In the first, GWAS was conducted on IDM (purple) and non-IDM (gray) SNP markers. For the second, GWAS was done using using the mean *M. glaziovii* allele dosages in 250Kb windows (*DoseGlaz*; dark red diamonds).

Five–fold Cross–validation Accuracy

Figure S15: Cross-validation accuracy for each Trait-Institute combination. Boxplots of accuracies from 5-fold cross-validation replicated 10 times per trait-institute-model. Three models were tested, the non-partitioned (ALL) model, the model without the IDM-component (nonIDMonly), and the partitioned model (PARTITIONED). In addition, two methods of partitioning the genome were compared for both the PARTITIONED and the nonIDMonly model: the IDM-based partition, and 15 different random partitions. Note that this means for each 15 random partition we did the full 5-fold x 10 reps, and we pool all random-partition accuracies within the relevant model as color-coded above. For each trait-institute dataset, we used the same 10 random partitions of the training-test for each model tested.

Figure S16: Genome-wide plots of significant introgression-trait associations. Manhattan plots summarizing genome-wide associations for 10 Trait-Institute analyses. Two mixed-linear model association analyses are shown, overlayed. In the first, GWAS was conducted on IDM (purple, top panel) and non-IDM (gray, middle panel) SNP markers. For the second, GWAS was conducted on 250Kb-window mean *M. glaziovii* allele dosages (Segment, orange, bottom panel). The size and alpha of *DoseGlaz* results are scaled with their -$\log_{10}$(p-value)). The horizontal lines represent the bonferroni-significance threshold for the *DoseGlaz* (purple line) and SNP GWAS (gray line).
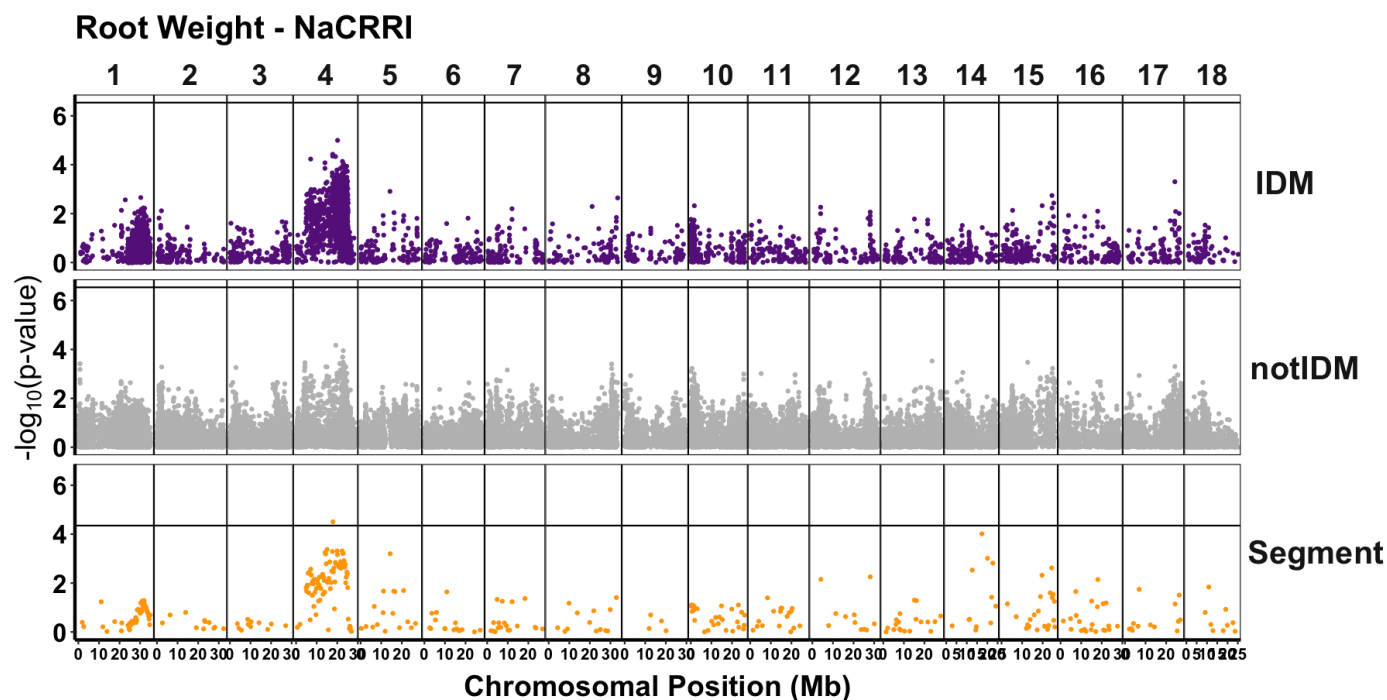
Figure S17: Genome-wide plots of significant introgression-trait associations. Manhattan plots summarizing genome-wide associations for 10 Trait-Institute analyses. Two mixed-linear model association analyses are shown, overlayed. In the first, GWAS was conducted on IDM (purple, top panel) and non-IDM (gray, middle panel) SNP markers. For the second, GWAS was conducted on 250Kb-window mean *M. glaziovii* allele dosages (Segment, orange, bottom panel). The size and alpha of *DoseGlaz* results are scaled with their -log$_{10}$(p-value)). The horizontal lines represent the bonferroni-significance threshold for the *DoseGlaz* (purple line) and SNP GWAS (gray line).
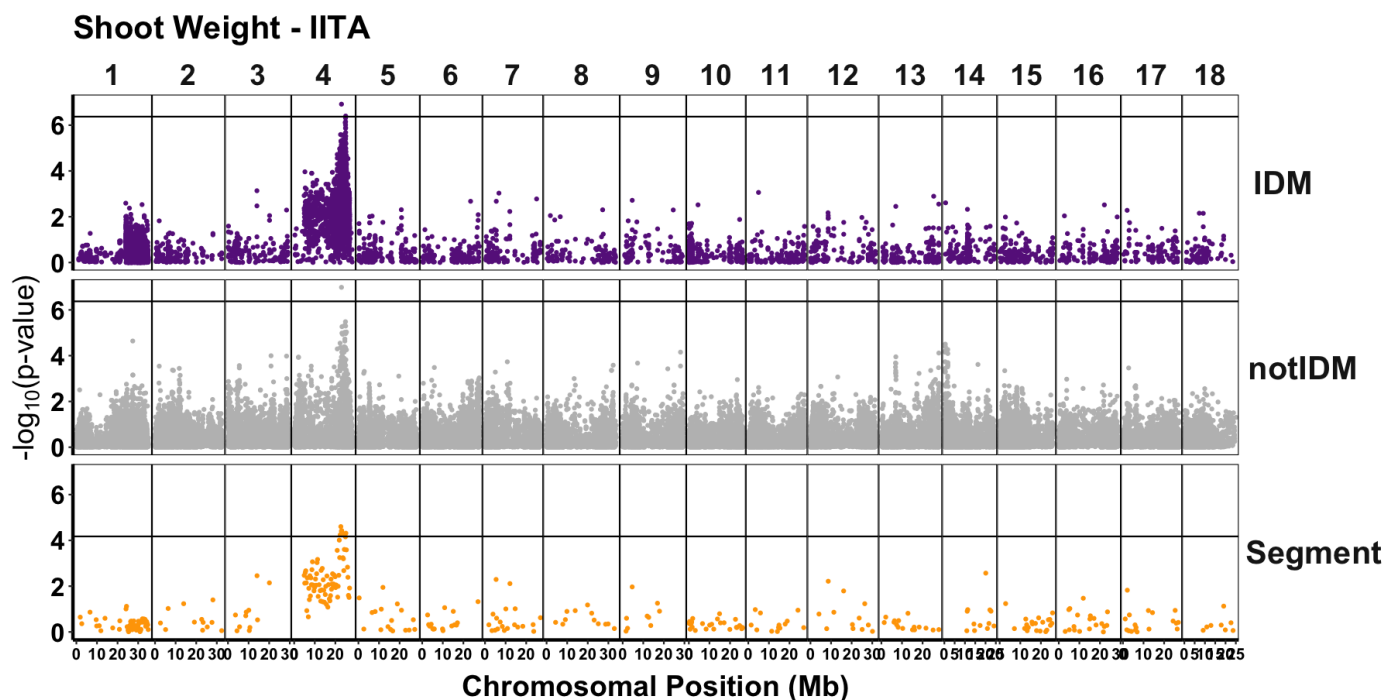
Figure S18: Genome-wide plots of significant introgression-trait associations. Manhattan plots summarizing genome-wide associations for 10 Trait-Institute analyses. Two mixed-linear model association analyses are shown, overlayed. In the first, GWAS was conducted on IDM (purple, top panel) and non-IDM (gray, middle panel) SNP markers. For the second, GWAS was conducted on 250Kb-window mean *M. glaziovii* allele dosages (Segment, orange, bottom panel). The size and alpha of *DoseGlaz* results are scaled with their -$\log_{10}$(p-value)). The horizontal lines represent the bonferroni-significance threshold for the *DoseGlaz* (purple line) and SNP GWAS (gray line).

Root Weight - NaCRRI

Figure S19: Genome-wide plots of significant introgression-trait associations. Manhattan plots summarizing genome-wide associations for 10 Trait-Institute analyses. Two mixed-linear model association analyses are shown, overlayed. In the first, GWAS was conducted on IDM (purple, top panel) and non-IDM (gray, middle panel) SNP markers. For the second, GWAS was conducted on 250Kb-window mean *M. glaziovii* allele dosages (Segment, orange, bottom panel). The size and alpha of *DoseGlaz* results are scaled with their -log$_{10}$(p-value)). The horizontal lines represent the bonferroni-significance threshold for the *DoseGlaz* (purple line) and SNP GWAS (gray line).
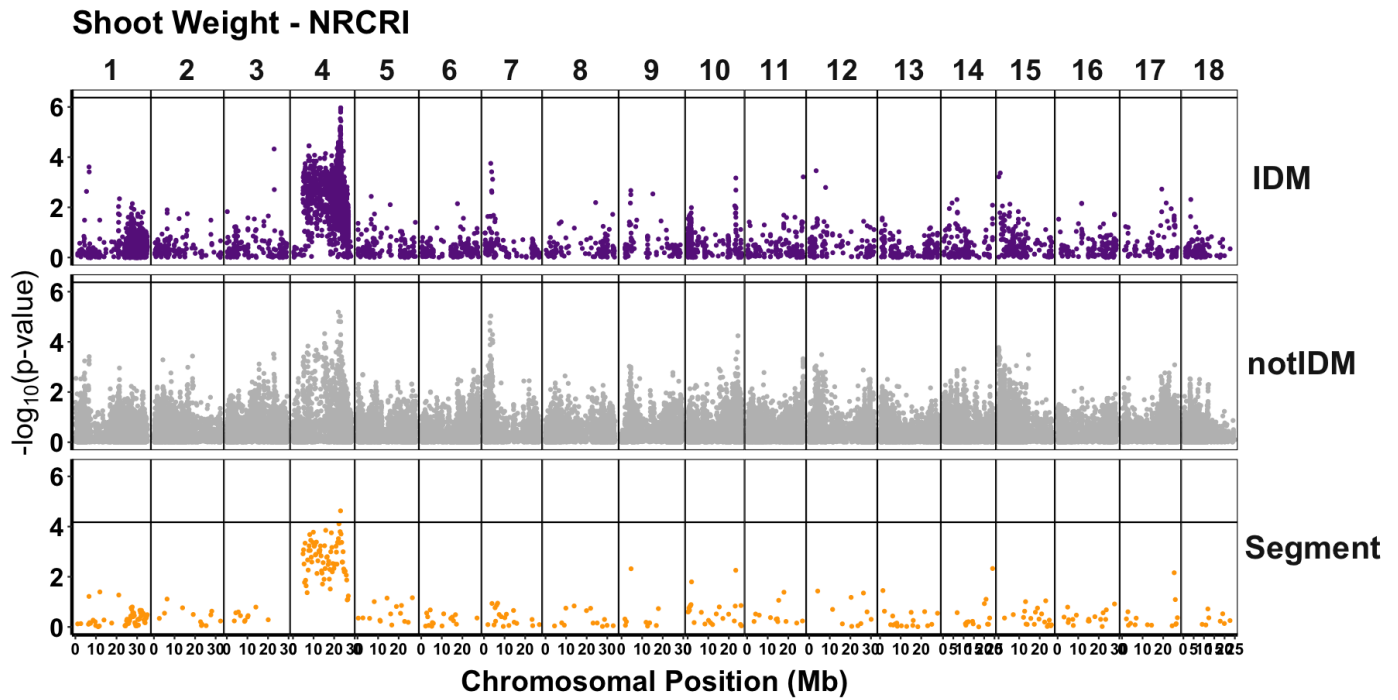
Figure S20: Genome-wide plots of significant introgression-trait associations. Manhattan plots summarizing genome-wide associations for 10 Trait-Institute analyses. Two mixed-linear model association analyses are shown, overlayed. In the first, GWAS was conducted on IDM (purple, top panel) and non-IDM (gray, middle panel) SNP markers. For the second, GWAS was conducted on 250Kb-window mean *M. glaziovii* allele dosages (Segment, orange, bottom panel). The size and alpha of *DoseGlaz* results are scaled with their -log$_{10}$(p-value)). The horizontal lines represent the bonferroni-significance threshold for the *DoseGlaz* (purple line) and SNP GWAS (gray line).
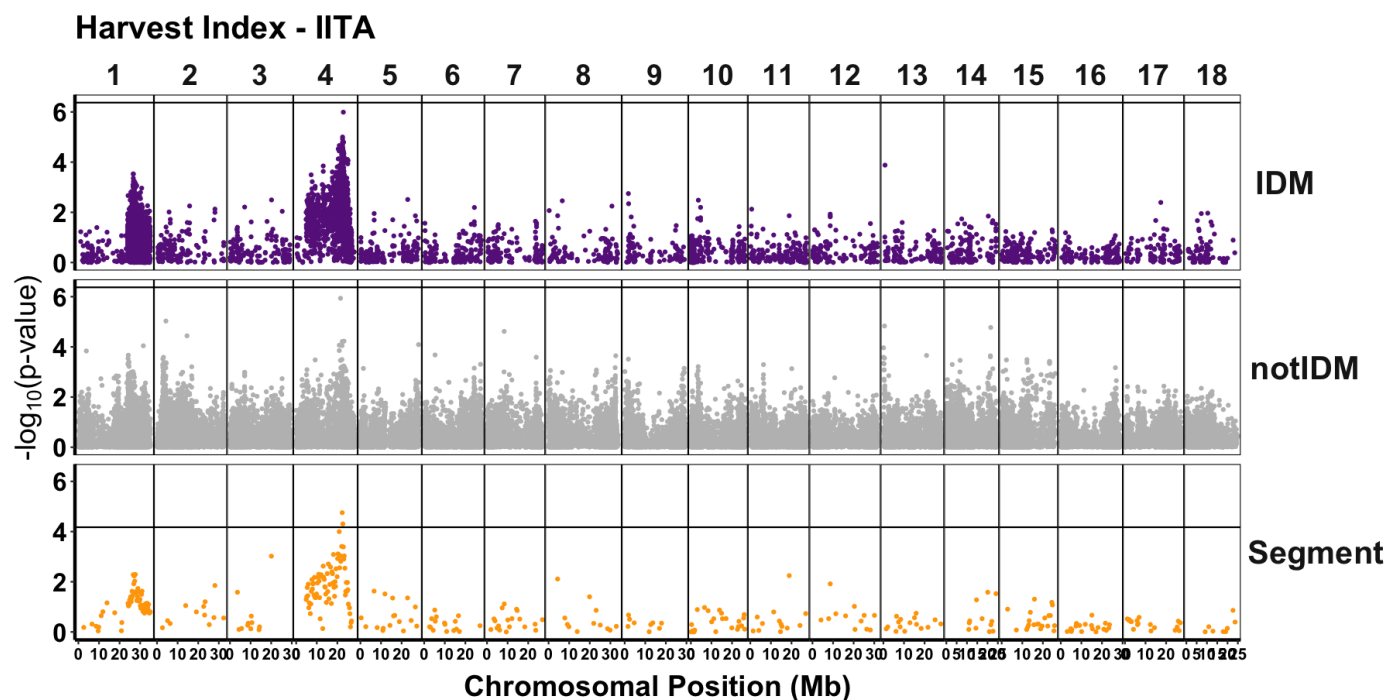
Figure S21: Genome-wide plots of significant introgression-trait associations. Manhattan plots summarizing genome-wide associations for 10 Trait-Institute analyses. Two mixed-linear model association analyses are shown, overlayed. In the first, GWAS was conducted on IDM (purple, top panel) and non-IDM (gray, middle panel) SNP markers. For the second, GWAS was conducted on 250Kb-window mean *M. glaziovii* allele dosages (Segment, orange, bottom panel). The size and alpha of *DoseGlaz* results are scaled with their -log$_{10}$(p-value)). The horizontal lines represent the bonferroni-significance threshold for the *DoseGlaz* (purple line) and SNP GWAS (gray line).
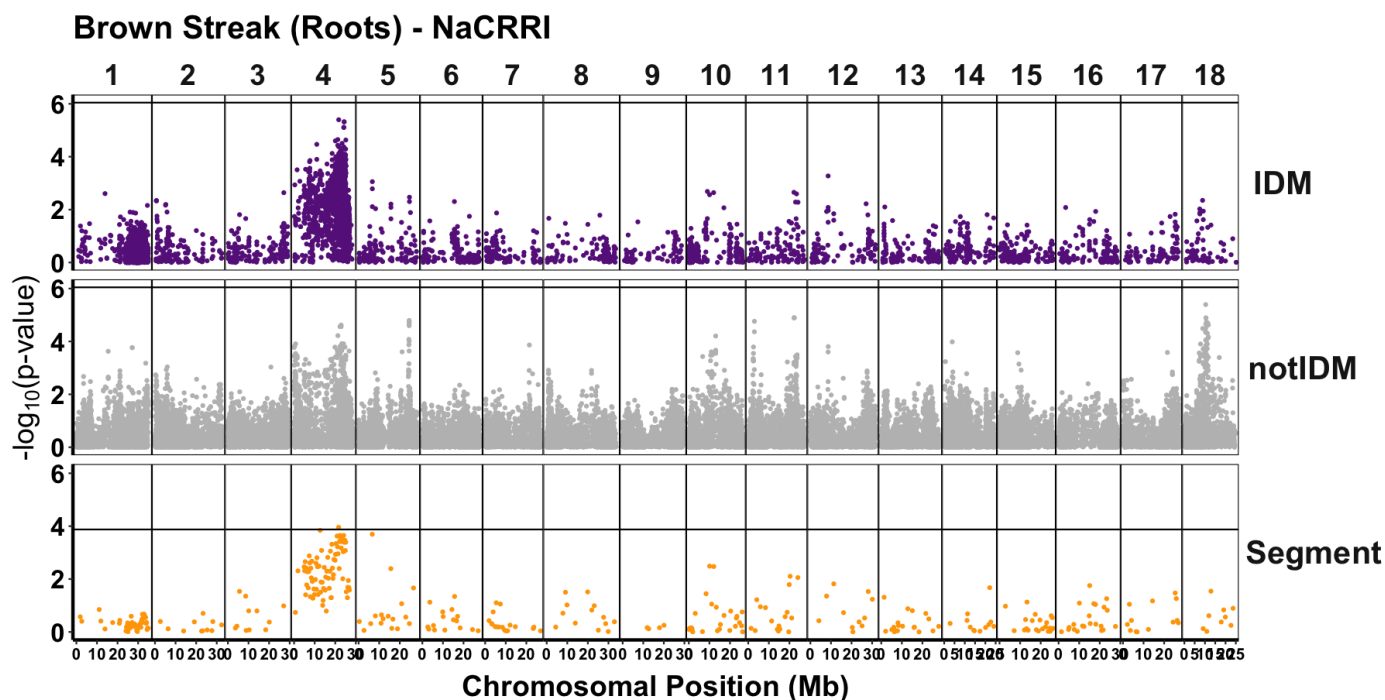
Figure S22: Genome-wide plots of significant introgression-trait associations. Manhattan plots summarizing genome-wide associations for 10 Trait-Institute analyses. Two mixed-linear model association analyses are shown, overlayed. In the first, GWAS was conducted on IDM (purple, top panel) and non-IDM (gray, middle panel) SNP markers. For the second, GWAS was conducted on 250Kb-window mean *M. glaziovii* allele dosages (Segment, orange, bottom panel). The size and alpha of *DoseGlaz* results are scaled with their -log$_{10}$(p-value)). The horizontal lines represent the bonferroni-significance threshold for the *DoseGlaz* (purple line) and SNP GWAS (gray line).
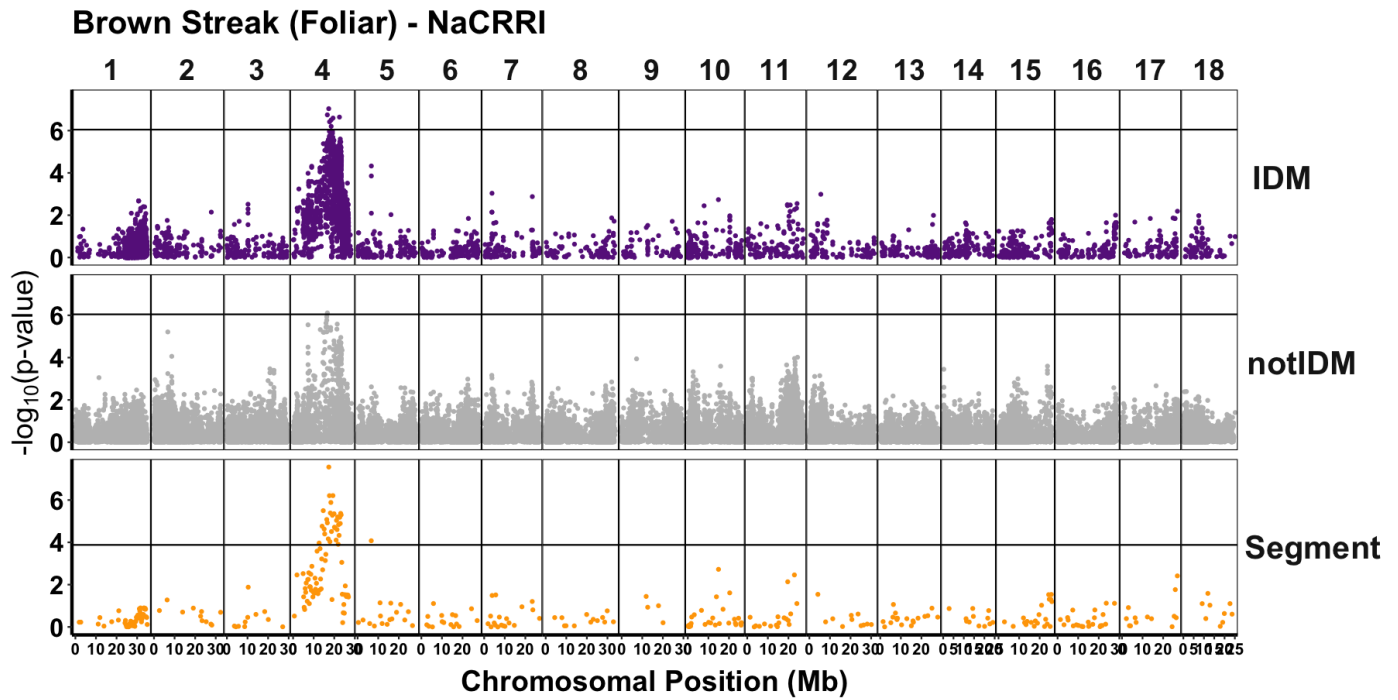
Figure S23: Genome-wide plots of significant introgression-trait associations. Manhattan plots summarizing genome-wide associations for 10 Trait-Institute analyses. Two mixed-linear model association analyses are shown, overlayed. In the first, GWAS was conducted on IDM (purple, top panel) and non-IDM (gray, middle panel) SNP markers. For the second, GWAS was conducted on 250Kb-window mean *M. glaziovii* allele dosages (Segment, orange, bottom panel). The size and alpha of *DoseGlaz* results are scaled with their -log_{10}(p-value)). The horizontal lines represent the bonferroni-significance threshold for the *DoseGlaz* (purple line) and SNP GWAS (gray line).

Figure S24: Genome-wide plots of significant introgression-trait associations. Manhattan plots summarizing genome-wide associations for 10 Trait-Institute analyses. Two mixed-linear model association analyses are shown, overlayed. In the first, GWAS was conducted on IDM (purple, top panel) and non-IDM (gray, middle panel) SNP markers. For the second, GWAS was conducted on 250Kb-window mean *M. glaziovii* allele dosages (Segment, orange, bottom panel). The size and alpha of *DoseGlaz* results are scaled with their -log$_{10}$(p-value)). The horizontal lines represent the bonferroni-significance threshold for the *DoseGlaz* (purple line) and SNP GWAS (gray line).

Figure S25: Genome-wide plots of significant introgression-trait associations. Manhattan plots summarizing genome-wide associations for 10 Trait-Institute analyses. Two mixed-linear model association analyses are shown, overlayed. In the first, GWAS was conducted on IDM (purple, top panel) and non-IDM (gray, middle panel) SNP markers. For the second, GWAS was conducted on 250Kb-window mean *M. glaziovii* allele dosages (Segment, orange, bottom panel). The size and alpha of *DoseGlaz* results are scaled with their -log$_{10}$(p-value)). The horizontal lines represent the bonferroni-significance threshold for the *DoseGlaz* (purple line) and SNP GWAS (gray line).