frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 09 January 2023
DOI 10.3389/fpls.2022.1041925

OPEN ACCESS

EDITED BY
Xiaofei Zhang,
International Center for Tropical Agriculture (CIAT), Colombia

REVIEWED BY
Tong Wei,
Beijing Genomics Institute (BGI), China
Zhiqiang Xia,
Hainan University, China

*CORRESPONDENCE
Evan M. Long
eml255@cornell.edu

SPECIALTY SECTION
This article was submitted to Plant Breeding, a section of the journal Frontiers in Plant Science

RECEIVED 11 September 2022
ACCEPTED 06 December 2022
PUBLISHED 09 January 2023

CITATION
Long EM, Romay MC, Ramstein G, Buckler ES and Robbins KR (2023) Utilizing evolutionary conservation to detect deleterious mutations and improve genomic prediction in cassava.
Front. Plant Sci. 13:1041925.
doi: 10.3389/fpls.2022.1041925

Utilizing evolutionary conservation to detect deleterious mutations and improve genomic prediction in cassava

Evan M. Long[1]*, M. Cinta Romay[2], Guillaume Ramstein[3], Edward S. Buckler[1,2,4] and Kelly R. Robbins[1]

[1]Plant Breeding and Genetics Section, School of Integrative Plant Science, Cornell University, Ithaca, NY, United States, [2]Institute for Genomic Diversity, Cornell University, Ithaca, NY, United States, [3]Center for Quantitative Genetics and Genomics, Aarhus University, Aarhus, Denmark, [4]United States Department of Agriculture-Agricultural Research Service, Robert W. Holley Center for Agriculture and Health, Ithaca, NY, United States

Introduction: Cassava (Manihot esculenta) is an annual root crop which provides the major source of calories for over half a billion people around the world. Since its domestication ~10,000 years ago, cassava has been largely clonally propagated through stem cuttings. Minimal sexual recombination has led to an accumulation of deleterious mutations made evident by heavy inbreeding depression.

Methods: To locate and characterize these deleterious mutations, and to measure selection pressure across the cassava genome, we aligned 52 related Euphorbiaceae and other related species representing millions of years of evolution. With single base-pair resolution of genetic conservation, we used protein structure models, amino acid impact, and evolutionary conservation across the Euphorbiaceae to estimate evolutionary constraint. With known deleterious mutations, we aimed to improve genomic evaluations of plant performance through genomic prediction. We first tested this hypothesis through simulation utilizing multi-kernel GBLUP to predict simulated phenotypes across separate populations of cassava.

Results: Simulations showed a sizable increase of prediction accuracy when incorporating functional variants in the model when the trait was determined by<100 quantitative trait loci (QTL). Utilizing deleterious mutations and functional weights informed through evolutionary conservation, we saw improvements in genomic prediction accuracy that were dependent on trait and prediction.

Conclusion: We showed the potential for using evolutionary information to track functional variation across the genome, in order to improve whole genome trait prediction. We anticipate that continued work to improve genotype accuracy and deleterious mutation assessment will lead to improved genomic assessments of cassava clones.

KEYWORDS

genetic load, deleterious mutation, cassava (*Manihot esculenta*), genomic prediction, evolutionary conservation

1 Introduction

Cassava (Manihot esculenta) is a root crop that is clonally propagated and grown widely in the tropical regions of Africa, Asia, and South America. It is estimated that cassava is a major caloric source for almost half a billion people around the world (Parmar et al., 2017; Ferguson et al., 2019). Although it is naturally an outcrossing perennial, it has been clonally propagated and grown as an annual since its domestication between 5,000-10,000 years ago (Wang et al., 2014). During the colonial era it was also brought to Africa, where today it is valued for its ability to grow with minimal inputs in marginally fertile lands.

Many generations of clonal propagation have caused cassava to accumulate genetic load that inhibits its potential crop performance. This genetic load is most apparent in the heavy inbreeding depression exhibited in cassava, as observed through low performance of selfed offspring (Rojas et al., 2009; de Freitas et al., 2016). Studies have shown that this genetic load is present as deleterious recessive mutations that are masked by heterozygosity which can be maintained through the clonal propagation (Ramu et al., 2017). With minimal sexual reproduction these deleterious mutations are maintained (McKey et al., 2010) and inhibit current breeding efforts to improve cassava performance (de Freitas et al., 2016).

Plant breeders have worked on various methods to detect and manage genetic load throughout history. Many crop species exist as polyploids, which enables them to more easily mask recessive deleterious mutations responsible for genetic load (van de Peer et al., 2021). Hybrid crop breeding has been another common method of applying strong selection pressures by selecting on inbred lines (Labroo et al., 2021), eliminating the possibility of recessive deleterious mutations. Some crops with similar high inbreeding depression to cassava, like potato, have made recent efforts to breed with inbred diploids (Bachem et al., 2019), however the deleterious mutations targeted by this methodology reduce plant viability.

During the past decade, plant breeders have seen the emergence of methodical application of genotyping and genomic selection as a method to improve breeding selections and leverage understanding of genomic information. Genomic selection, which uses genome markers and a phenotyped training population to predict unobserved offspring performance, can decrease selection cycle time and improve selection accuracy. Efforts have been made to improve genomic selection by using causative knowledge, however understanding the true causative elements in the genome is not a trivial exercise. Many studies have shown that benefits from including genome-wide association (GWA) hits in genomic prediction can diminish when predicting unrelated material (Cheruiyot et al., 2022), indicating population specific quantitative trait locus (QTL) or a misinterpretation of a variant as causative, when it is only in high linkage disequilibrium (LD) with the causative variant (Cheruiyot et al., 2022). For cassava, an ideal genomic annotation would explain underlying causative elements, while being consistent across populations structures.

Regarding genetic load, evolutionary conservation has shown to be an effective method to assess deleterious mutations and explain functional variation (Xiang et al., 2019) in a population agnostic manner. Multiple studies in crops such as maize (Yang et al., 2016; Ramstein and Buckler, 2022), sorghum (Valluru et al., 2019; Lozano et al., 2021), and barley (Kono et al., 2019) have demonstrated potential benefits for detecting and using deleterious mutations in genomic prediction. The potential benefit of understanding these deleterious mutations in cassava will be limited by the absolute number of mutations and how much variation of agronomic traits they each explain.

The purpose of this study is first, to identify likely deleterious mutations in cassava, and second to evaluate their potential impact on genomic prediction for the goal of improving future breeding selections. We sequenced, assembled, and gathered 52 genomes from species that all shared ancestry within the last 50 million years in order to score conservation and detect deleterious mutations.

We designed an experiment that uses evolutionary information to augment genomic predictions within and across two different populations of 1048 cassava clones present in two different breeding programs in Sub-Saharan Africa, the International Institute of Tropical Agriculture (IITA), Ibadan, Nigeria, and the National Crops Resources Research Institute (NaCRRI), Namulonge, Uganda. By performing phenotype simulations using real genotypic data and generating genomic predictions with known, simulated QTL, we first evaluated the best possible benefit of including causative information in our genomic predictions under different scenarios. We then used genomic and phenotypic data from these cassava clones to test genomic predictions, while including various functional annotations based on deleterious mutations.

2 Results

2.1 Evolutionary conservation

Utilizing many germplasm resources, we sampled, sequenced and assembled 27 Euphorbiaceae species (Supplementary Table 1). These assemblies were combined with available genome from Euphorbiaceae and other related species to form a set of 53 species, including cassava. We obtained multiple sequence alignments from for each gene, requiring transcript alignment of ≥90% of length of the cassava gene. Only the best matching ortholog from each species was retained and, of the ~26k genes examined, 24565 genes had ≥4 orthologs, allowing them to be scored for evolutionary conservation using PAML's baseml tool. Over

half of all base pairs across these genes have an alignment depth of ≥31 species (Figure 1). The large number of aligned orthologs from the many species to measure conservation is benefited from sampling species from within shorter evolutionary time, although it is limited by poorer gene reconstruction in assemblies from short-read sequence.

2.2 Deleterious mutations

We used evolutionary conservation and predicted protein mutation effects to classify the deleterious effects of 66k nonsynonymous SNPs segregating in the two target populations. Firstly, we used the intersection of baseml evolutionary rate and SIFT deleterious scores to classify 2,210 deleterious sites that are segregating in both cassava populations (Figure 2). While both methods rely on evolutionary information, the high coincidence of low evolutionary rate and low SIFT score support their signal for functionally important sites in the genome. Deleterious burden for each clone was then calculated as the number of derived alleles at these sites. We separated this deleterious burden into homozygous and heterozygous genetic load. Genome wide association for all nonsynonymous sites as well as the deleterious sites was performed on fresh root yield and dry matter percentage traits, and some loci passed Bonferroni significance testing for fresh root yield (Supplementary Figures 4, 5). Secondly, we leveraged a RandomForest prediction model to weight the functional importance of the nonsynonymous mutations. This prediction produces a score between 0-1, a quantitative weight for the functional importance of each amino acid residue altered by mutations at the nonsynonymous sites (Figure 3).

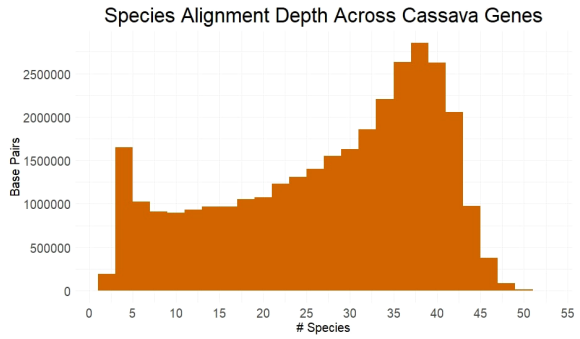

FIGURE 1
Species Alignment Depth Across Cassava Genes. Alignment depth represented by the number of species with homologous alleles in each multiple sequence alignment at any given protein coding base pair in the cassava genome.

2.3 Phenotype simulation

To validate our methodology and guide our expectations we performed genomic predictions using simulated phenotypes on 1048 cassava clones originating from IITA and NaCRRI breeding programs. These simulations represent some best-case scenarios for genomic prediction, where all QTL and their effect sizes are known.

The simulated QTL effects represent a suite of different genetic architectures ranging from highly complex genetic traits controlled by thousands of small effect QTL to oligogenic traits controlled by a handful of large effect QTL. These genetic architectures are represented by the proportion of the 66k variants simulated as causative QTL (Figure 4). These 66k variant sites were selected using nonsynonymous sites that showed high conservation (low evolution rate) from baseml. We modeled a range of dominance levels at each QTL in order to match our empirical scenario more closely in cassava (Supplementary Figure 1), where genetic load due to recessive deleterious alleles are expected to affect many agronomic, fitness related, traits (Bosse et al., 2019).

2.4 Genomic prediction with simulated phenotypes

Once QTL effects were modeled, we then calculated phenotypes for each of the 1048 clones (Supplementary Figure 2), where a positive effect is attributed to the ancestral allele. To Evaluate the effect of QTL structure, prediction model, and population, we performed genomic predictions. For all predictions in this study, we performed cross-population and within-population predictions designated as follows: IITA cross-validation (IITA_CV), NaCRRI cross-validation (NaCRRI_CV), Training with the IITA population and predicting in the NaCRRI population (IITA->NaCRRI), and "Training with the NaCRRI population and predicting in the IITA population (IITA->NaCRRI). Cross-population prediction accuracy is calculated by masking all phenotypes in one population and predicting using the other, then calculating the correlation between the true phenotype and the predicted phenotype. Within-population prediction accuracy is calculated similarly, using a 10-Fold prediction scheme where phenotypes in 10% of a population are masked and predicted by the other 90%.

We saw a marked increase in prediction accuracy when including the QTL information into the prediction model only when the trait was controlled by less than around 100 QTL (Figures 5C, D). Complex traits that are controlled by many small effect QTL across the genome show no increase in prediction accuracy with the inclusion of causative information (Figures 5A, B). For traits with an intermediate number of QTL (Figure 5C), the improvements in prediction accuracy are further increased by weighting the QTL information by their relative effect sizes. While the improvements are visible in both cross-population

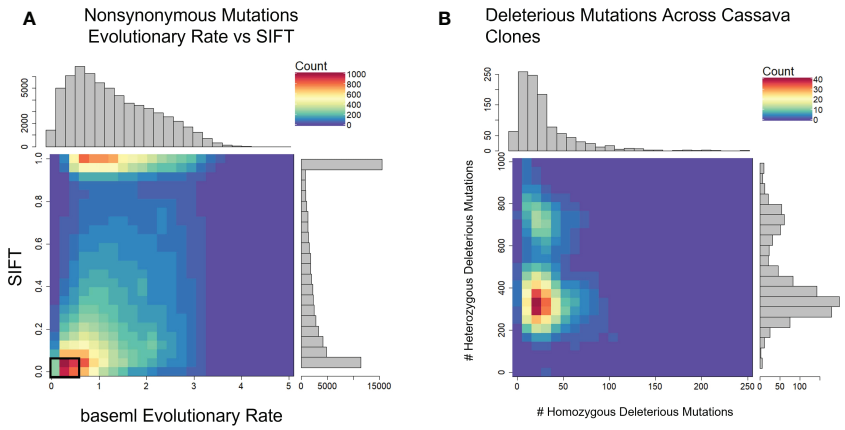

FIGURE 2
Defining Deleterious Mutations. **(A)** baseml evolutionary rate is plotted against SIFT scores. Deleterious mutations were classified as derived alleles at those sites with a baseml evolutionary rate < 0.5 and a SIFT score < 0.05 (Black box). **(B)** Distribution of homozygous and heterozygous deleterious mutations across 1048 cassava clones.

and within- population predictions, the improvements show some evidence of being more pronounced in cross-populations scenarios. These simulations show that even with perfect knowledge of QTL effects, improvements in prediction accuracy from using this information are limited by the relative abundance of those QTL.

2.5 Genomic prediction utilizing functional annotation

With deleterious mutations and functional weights for the segregating nonsynonymous sites, we mirrored the genomic predictions that we previously performed using simulated phenotypes, only this time using real data collected on the 1048 cassava clones.

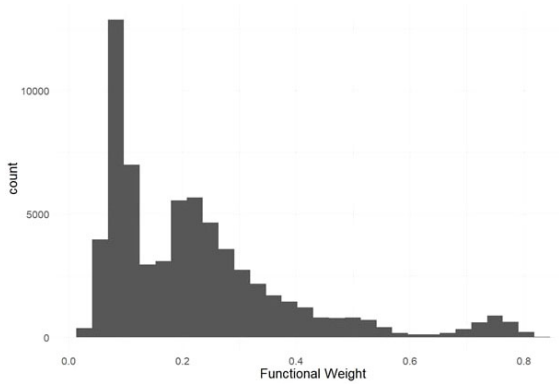

FIGURE 3
Predicted Functional Weights. Histogram of functional weights produced through RandomForest prediction of conservation for nonsynonymous variant sites. High functional weights correspond to highly conserved sites where nonsynonymous mutations are predicted to have large functional effects.

We predicted two different traits common in cassava breeding trials, fresh root yield and dry matter percentage, using the same cross-population and within-population scenarios previously shown. Multiple genomic prediction models were tested to evaluate the value of including the functional annotations.

Our two examples of a baseline prediction, where no functional information is present, are genomic prediction using the input marker data set and a genome-wide imputed dataset. In predicting fresh root yield, our results show that imputation alone does not improve cross-population prediction accuracy, however it does show some positive effect on within-population prediction (Figure 6). However, when including only imputed, segregating, non-synonymous variants, the prediction accuracy in cross-population predictions does increase over the two baseline models. Finally, we observed a further increase in prediction accuracy when weighting the non-synonymous variants and including derived genetic load from the deleterious mutations for both the cross-population predictions of fresh root yield and for within-population predictions in among the NaCRRI clones (Figure 6; Supplmentary Figure 6). For genomic prediction of cassava tuber dry matter percentage, we observed mostly negative or neutral effects of imputation and inclusion of deleterious annotations (Figure 7; Supplementary Figure 7). The improvements from functional information in predicting fresh root yield suggest it is correlated with fitness signals captured by the evolutionary information, while dry matter percentage may represent different, historical selection pressures.

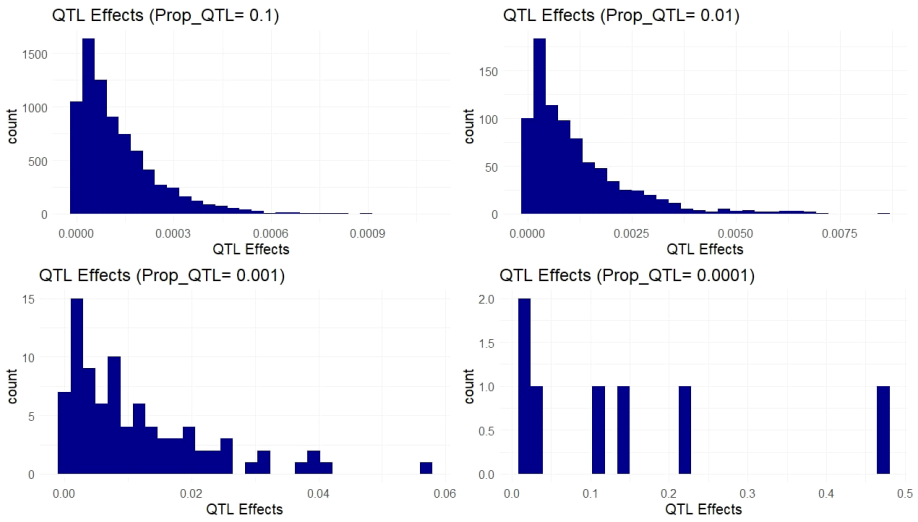

FIGURE 4
Simulated QTL Effects. Histograms show count of QTL effects in one example simulation. Each facet shows a genetic architecture with different proportions of the markers acting as QTL (resulting in ~ 6600, 660, 66, and 6 QTL on average). The x-axis represents the positive effect of carrying the ancestral allele at a given QTL.

3 Discussion

Genetic load, as defined as the accumulation of deleterious mutations through domestication, drift, mutation-selection balance and other means, has been identified as an impediment to the genetic value of a crop (Agrawal and Whitlock, 2012; Smýkal et al., 2018). Through simulation, we explored the possible scenarios in which knowing the exact deleterious mutations could improve breeding selections. In this study, we went on to use evolutionary conservation and genomic information to quantify deleterious mutations in cassava clones, as well as predict their potential effects.

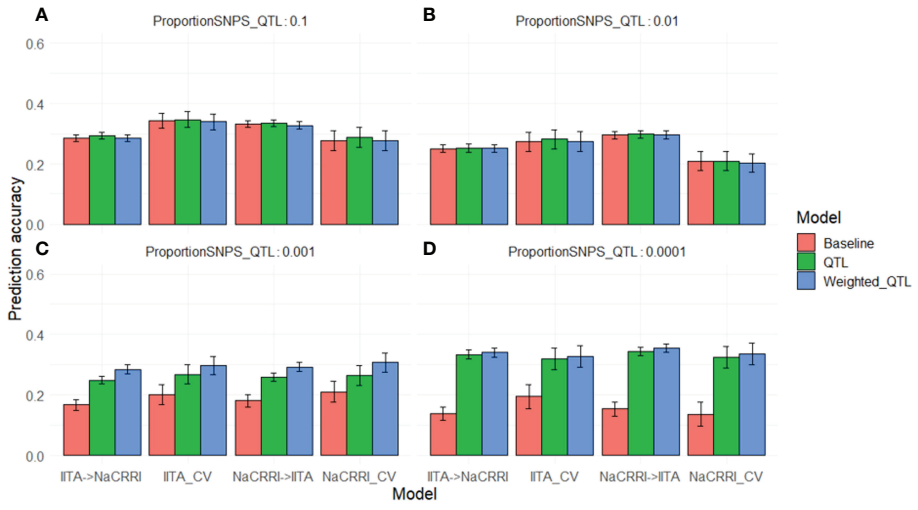

FIGURE 5
Genomic Prediction Accuracies with Simulated QTL. Prediction accuracies are shown on the y-axis as the correlation between predicted andtrue breeding values. The x-axis delineates the prediction scenario being tested. Barplot color corresponds to the genomic information used in the prediction model. Error bars represent a 95% confidence interval for simulations. Simulations were repeated with different proportions of the markers acting as causative QTL: 0.1 **(A)**, 0.01 **(B)**, 0.001 **(C)**, and 0.0001 **(D)**.

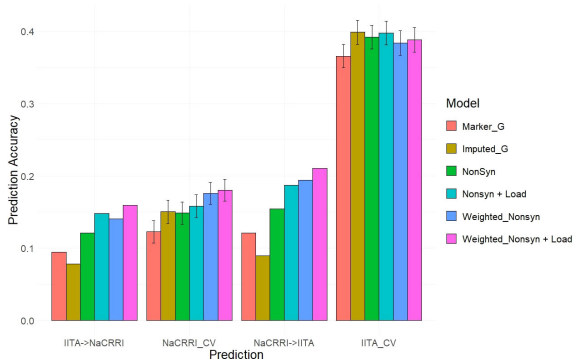

FIGURE 6
Fresh Root Yield Genomic Prediction Leveraging Deleterious Annotations. Prediction accuracy is measured in cross-population and within-population prediction scenarios. Genomic models are represented as bar graph colors where various genomic and deleterious data are used in the genomic prediction. Error bars represent a 95% confidence interval for within-population 10-fold prediction.

3.1 Simulation informs genomic prediction potential

The simulation of phenotypes under differing genetic architectures allowed us to manage expectations for the best possible scenarios in which understanding the causative variation of a trait could help inform genomic selection decisions. As we only observed benefits to genomic prediction under scenarios with<~100 QTL, it is clear that LD structure captured by genome wide markers is sufficient for genomic prediction under highly complex genetic architectures (Figure 5). The scenario with the fewest QTL (<~10) represents a more Mendelian or oligogenic architecture, which might benefit more from a marker assisted selection methodology, but it follows that traits with higher effect sizes of QTL will see more improvements from causative knowledge in genomic prediction. Interestingly, within-population predictions showed smaller, but still substantial benefits in genomic prediction accuracy. These results indicate that our empirical predictions have the potential to benefit from deleterious mutation annotations, only if there are a few or intermediate number of QTL (<~100) with substantial effects. Importantly, the expected benefits shown through simulations depend directly upon the population and LD structure in our tested clones and cannot necessarily be useful to interpret potential benefits in other scenarios.

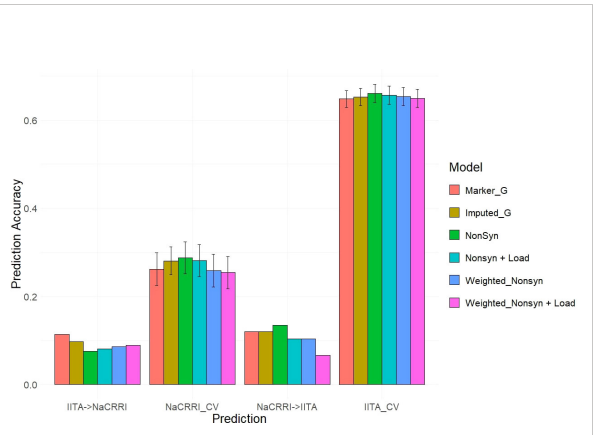

FIGURE 7
Dry Matter Percentage Genomic Prediction Leveraging Deleterious Annotations. Prediction accuracy is measured in cross-population and within-population prediction scenarios. Genomic models are represented as bar graph colors where various genomic and deleterious data are used in the genomic prediction. Error bars represent a 95% confidence interval for within-population 10-fold prediction.

3.2 Evolution conservation reveals deleterious mutations

We used evolutionary conservation and protein annotations to classify certain mutations as deleterious. By aligning over 50 species of relatively recent ancestry, we were able to assess the conservation status of a large majority of the cassava genome. We used separate neutral trees for each gene, rather than the entire chromosome or species, to address the difference between gene ancestry common in plants due to historical gene and genome duplication. Because of the millions of years of evolution, it is very difficult to predict the sizes of selection coefficients from evolutionary conservation alone (Huber et al., 2020). We then needed predicted protein effects of these mutations from SIFT to refine our set of putatively deleterious mutations. After defining our deleterious alleles, we separated the assessment of deleterious load into homozygous and heterozygous, because most deleterious mutations are assumed to be recessive (Bosse et al., 2019) and cassava has been shown to mask deleterious mutations through heterozygosity (Ramu et al., 2017). These assessments of genetic load are at least partially validated by a negative correlation (R=-0.18) between plant yield and homozygous deleterious mutations (Supplementary Figure 3).

As previously mentioned, evolutionary conservation alone cannot easily resolve effect sizes of mutations. For this reason, we used protein perturbation information from SIFT and UniRep to prioritize functional variants similar to work recently done in Maize (Ramstein and Buckler, 2022). Another advantage of this weighting method is that it does not imply a directional effect of the mutations, thereby allowing for potential positive or adaptive effects (Loewe and Hill, 2010) of derived mutations at conserved sites. While most derived alleles at conserved positions are predicted to be deleterious, these derived alleles could

represent directed selection from domestication or adaptive evolution specific to cassava.

3.3 Leveraging functional data in genomic prediction

The inclusion of deleterious and functional mutations derived from evolutionary conservation showed promising value in informing the genetic value of cassava clones. Our results displayed improvements for cross-population predictions of fresh root yield as well as some of the within-population predictions in NaCRRI (Figure 6). This follows with the understanding that total plant growth, and even root yield, are correlated with total plant fitness (Pan and Price, 2001), while root dry matter percentage, which is primarily a quality trait, likely has little direct correlation with evolutionary fitness (Figure 7). We expect this trend would continue for other traits; however, few traits are measured identically across multiple populations.

In this study, we used multi-kernel GBLUP methods of genomic prediction to partition the additive and dominant genetic effects, while substituting unweighted and weighted genomic relationship matrices formed from subsets of the genomic data. These methodologies rely on the assumption that our selected functional variants, and the weights prescribed to them, are derived from a separate, and more functional, distribution of effects from a default, genome-wide relationship. Other methods, including Bayesian models, exist to prioritize functional information in genomic prediction, however multiple studies have found it to be difficult to prescribe consistent, significant differences in prediction accuracy results between them and GBLUP models, and the specific benefit of one method or the other are often situational (Moghaddar et al., 2019; Khansefid et al., 2020; Cheruiyot et al., 2022).

3.4 Reflections on load

In an effort to improve cassava's role as a reliable food source around the world, our results show the importance and potential of addressing the impact of genetic load. We used evolution and protein annotations to determine these deleterious mutations responsible for genetic load. It is important to note that, while the methods used in this study detected impactful deleterious variation across the genome, they ignore the many deleterious mutations likely found in regulatory regions of the genome.

The improvements made in genomic prediction validate the effects of these deleterious mutations and offer one possible avenue for their potential application. As observed in the within-population prediction of IITA, where prediction accuracy is higher and unaffected by our annotations, the application of this understanding of genetic load may not be beneficial in every breeding scenario, however cross-population prediction is not the only instance where deleterious information may prove informative. Rapid cycle recurrent selection, where generations of selection occur without phenotyping, could be another situation in which tracking functional information across the genome could improve genomic selection decisions. As generations of selection occur, linkage disequilibrium between causative mutations and genome-wide markers breaks down, making the functional tracking of causative effects more impactful in prediction.

In addition to genomic prediction scenarios, the understanding of the deleterious mutations responsible for genetic load in cassava could suggest alternative methods for crop improvement. Many crops today utilize hybrid breeding, where multiple groups of inbred parents are bred for use in creating a superior hybrid. Selecting on inbred individuals exposes recessive, or partially recessive, deleterious mutations, allowing them to be effectively purged in fewer generations. While difficulties due to severe inbreeding depression in cassava have hindered this genre of breeding, efforts being made in crops like potato show it's potential in a crop burdened by heavy genetic load (Bachem et al., 2019). Doubled haploidization has been a common tool in some inbred crops, while historically difficult to implement in some crops like cassava, however newer implementations such as those reported from ScreenSys (https://www.screensys.eu) offer a possible method of producing enough viable embryos for crops with heavy inbreeding depression like cassava. (Nasti and Voytas, 2021). With the understanding of the extent to which deleterious mutations account for missed potential in cassava performance, further consideration for how to effectively purge genetic load will be needed.

Historical evolution and population genetics continues to shed light on our understanding of genomic functions, as seen in our study in cassava. We showed the utility of using evolutionary derived deleterious mutations to improve genomic prediction across cassava populations. Additionally, the genetic load was identified from<~100 homozygous deleterious mutations per clone (Figure 2). This number of mutations could be the target of further improvement through gene editing or other means. In the future, as genome sequencing accelerates, coupled with our understanding of protein functions, we may be able to make targeted decisions to purge genetic load from cassava and advance genetic gains.

4 Methods

4.1 Euphorbiaceae sequencing & assembly

We gathered a total of 52 related species, 26 of which we sequenced and assembled, to evaluate evolutionary conservation

across the cassava genome. In order to maximize the amount of evolutionary time sampled, while maintaining reliable alignments to cassava, we sampled 26 species across the Euphorbiaceae family, to which cassava belongs. We then sequenced these species using Illumina NovaSeq-6000. Genome sizes were estimated using k-mer spectra in order to estimate sequence input coverage for assembly (https://bioinformatics.uconn.edu/genome-size-estimation-tutorial/). Additional short-read sequences were downloaded from SRA corresponding to 11 unspecified Euphorbiaceae taxa (Liu et al., 2019). We then used a short-read sequence assembler MEGAHIT (Li et al.,), with modified parameters of "-m 0.2 -t 10 –no-mercy –min-count 3 –k-min 31 –k-step 20" to create contig assemblies. We additionally obtained long-read sequences using PacBio Sequel II for 7 species among our sampled Euphorbiaceae taxa. These sequences were assembled using Hifiasm (<xr rid="r6">Cheng et al., 2021</xr>) utilizing default settings. An additional 15 genome assemblies from other related species were downloaded from SRA and added to our assembled genomes resulting in a total of 52, excluding cassava (Supplementary Table 1).

4.2 Sequence alignment and evolutionary conservation

We used gene alignments from Cassava V7.1 gene annotations to the 52 species to extract homologous gene sequences for multiple sequence alignment. Gene transcripts were aligned using minimap2, and the best aligned region with >= 90% alignment length matching was retained as homologous coding sequences for each species were then extracted and aligned using MAFFT (Katoh et al., 2002) multiple sequence alignment. With a multiple sequence alignment for each gene, we then generated gene trees using RAxML (Stamatakis, 2014), and calculated evolutionary rates using baseml from the PAML (Yang, 2007) suite of tools. We then identified ancestral alleles at every site across the genic regions of the genome, using the ancestral node containing Manihot, Hevea, and Cnidoscolus genera. We used evolutionary conservation to select representative gene models for each gene, as well as only retaining genes with 5' and 3' untranslated regions annotated resulting in ~25k genes models.

4.3 Deleterious mutations

We used evolutionary conservations & protein structure conservation to identify deleterious mutations and produce weights for functional importance of sites across the cassava Genome. Deleterious mutations were categorized as sites with a baseml evolutionary rate of <0.5 and a "Sorting Intolerant From Tolerant" (SIFT) score of < 0.05 (Ng and Henikoff, 2003). Additionally, we required deleterious sites to have < 20% minor allele frequency in the cassava HapMap (Ramu et al., 2017) (Figure 2).

In addition to identifying a binary classification of deleterious, we used a RandomForest model to obtain a quantitative prediction of conservation similar to a previously reported method reported (Ramstein and Buckler, 2022). We used baseml evolutionary rates to classify nonsynonymous sites as either conserved (evolutionary rate< 0.3) or non-conserved (evolutionary rate > 2), while sites with values outside these ranges were excluded from model training. SIFT, UniRep, and 100bp windowed GC% totaling ~500 predictors in the RandomForest model implemented by the R package "ranger" (Wright and Ziegler, 2017). From the SIFT database, we used both the mutation type and SIFT score, which gives the predicted deleterious effect of a base-pair substitution. UniRep is a deep learning technique which characterizes protein structure (Alley et al., 2019), which we used to produce 256-unit representations of each protein and its associated mutated forms (https://github.com/churchlab/UniRep).

To increase the number of observations in the model, we used both the known HapMap mutations and *in silico* non-synonymous mutations at every possible site in our gene models. This resulted in over 1 million non-synonymous mutations whose genomic conservation could be modeled. We then used a leave-one-out prediction scheme where each of the 18 cassava chromosomes was left out of model training and predicted by the other 17. This method produced a predicted value between 0-1 for each of the ~66k nonsynonymous, segregating mutations used in this study (Figure 3).

4.4 Phenotypic & genotypic data

Phenotypic and genotypic data for 1048 cassava clones were downloaded from *cassavabase.org* representing two populations of breeding lines. The first population is from a breeding program at International Institute of Tropical Agriculture (IITA) in Nigeria, while the second is from a breeding program at National Crops Resources Research Institute National Crops Resources Research Institute (NaCRRI) in Uganda, representing breeding material for West and East Africa, respectively. Genotypes for the associated clones were downloaded from the "*East Africa Clones Dart-GBS 2020*" genotyping protocol on cassavabase.org containing 23,431 variants. Plant phenotypes for fresh root yield and dry matter percentage were downloaded from *cassavabase.org* and prepared according to previously described methods (https://wolfemd.github.io/GenomicSelectionManual/index.html).

We then performed genotype imputation using the cassava haplotype map using Beagle5 (Browning et al., 2018), with an **Ne=100,** resulting in ~26M variants. These variants were then filtered down to two genome-wide marker sets, one being a

thinned sample of ~135k genome-wide SNPs, and the other being all non-synonymous sites segregating in both populations resulting in ~66k genome-wide variants. The input marker genotypes, the imputed sample, and the imputed non-synonymous sites will be used in genomic prediction analyses.

4.5 Causative variation simulation

We used quantitative trait loci (QTL) simulation, replicated 50 times, to model the potential benefits of knowing causative variants in genomic prediction. This simulation begins by sampling QTL across the 66K variant sites from a binomial distribution with the probability of being a QTL varied across possible values of 10^{-1}, 10^{-2}, 10^{-3}, and 10^{-4}. The effect sizes for these QTL were then sampled from a gamma distribution using the *rgamma* function in R, with the shape parameter=1, with the ancestral allele set as having a positive effect. Lastly a dominance effect for each QTL was sampled from normal distribution "rnorm(mean = 2,sd=0.3)", restricting to dominance<=2 (Supplementary Figure S1). Phenotypes were then generated for the 1048 cassava clones. Residuals were then simulated such that the trait had a heritability of approximately 0.3.

We performed cross-population and 10-Fold within-population predictions using the simulated data, with and without QTL information incorporated into the prediction model. Genomic prediction was performed by using GBLUP methods fit using ASReml, with additive and dominance effects modeled as separate kernels. For all models described, residuals are represented by ε and modeled as random with $\boldsymbol{\varepsilon}\sim N(\mathbf{0}, \mathbf{I}\,\sigma_\varepsilon^2)$.

For prediction using simulated phenotypes, we compared three different models. The first model represents our baseline prediction:

$$y = 1\mu + Z_A a + Z_D d + \varepsilon$$

Where y is the simulated phenotype, μ is the phenotype mean, **a** is the vector of additive genetic effects, $\mathbf{Z_A}$is the incidence matrix, and $\mathbf{a}\sim N(\mathbf{0},\mathbf{G_A}\ \sigma_a^2)$, G_A is an additive genomic relationship matrix produced using the VanRaden (VanRaden, 2008) method, and σ_a^2 is the additive genetic variance.

$$G_A = \frac{MM'}{\sum_i^n (2p_i*(1-p_i))}$$

Where M is the centered genotype matrix (where genotypes are stored as dosages of 0,1, and 2 referring to being homozygous for reference allele, heterozygous, and homozygous for the alternate allele, respectively) and p_i is and allele frequency at the i^{th} locus. $\mathbf{Z_D}$ and **d** are analogous to the additive method, with the exception that a dominance genomic relationship matrix is produced using the Nishio and Satoh (Nishio and Satoh, 2014) method.

$$G_D = \frac{DD'}{\sum_i^n (2p_i*(1-p_i))^2}$$

Where the entries of D are given as $-2p_i^2$ for the homozygous reference allele, $2p_i*(1-p_i)$ for the heterozygote, and $2(1-p_i)^2$ for the homozygous alternate allele.

The second model includes additive and dominance QTL relationship matrices formed in identical manner to the $\mathbf{G_A}$ & $\mathbf{G_d}$ matrices, but only utilizing the known QTL sites in the genomic relationship matrices:

$$y = 1\mu + Z_{AQTL}aQTL + Z_{DQTL}dQTL + \varepsilon$$

The final model includes weighted QTL matrices based on their effect size:

$$y = 1\mu + Z_{AW}aw + Z_{DW}dw + \varepsilon$$

Here the weighted matrices are formed using modified methods of the previously cited methods. The weighted additive matrix given by:

$$G_{AW} = \frac{MWM'}{\sum_i^n (2p_i*(1-p_i)*w_i)}$$

Where M is the scaled genotype matrix. W is a diagonal matrix with w_i along the diagonal, w_i and p_i are the weight and frequency for the i^{th} locus, respectively.

The weighted dominance matrix is modified in a similar fashion to the additive matrix:

$$G_{DW} = \frac{DWD'}{\sum_i^n (2p_i*(1-p_i)*w_i)^2}$$

Where the entries of D are given as $-2p_i^2$ for the homozygous reference allele, $2p_i*(1-p_i)$ for the heterozygote, and $2(1-p_i)^2$ for the homozygous alternate allele.

4.6 Genomic prediction models in empirical data

The genomic prediction models used for real breeding program phenotypes follow a similar pattern to our simulated scenario, with a few notable differences.

First, our ground truth for the phenotype of each clone was the best linear unbiased estimate (BLUE) using a model like those previously used in cassava plot level traits (Wolfe et al., 2017) and those suggested for use with African cassava breeding data (https://wolfemd.github.io/GenomicSelectionManual/index.html):

$$y = X\beta + Z_{block(rep)}b + Z_{rep(trial)}t + \varepsilon$$

where y is the vector of the phenotype, β included a vector of fixed effects for the population mean, the location–year

combination, the number of plants harvested per plot, and germplasm ID with design matrix **X**. Replications were nested in trials, treated as random, and represented by the design matrix $\mathbf{Z_{rep(trial)}}$ and the effects vector $\mathbf{t}\sim N(\mathbf{0},\mathbf{I}\,\sigma_t^2)$. Blocks were nested in replications, treated as random, and represented by the design matrix $\mathbf{Z_{block(rep)}}$ and the effects vector $\mathbf{b}\sim N(\mathbf{0},\mathbf{I}\,\sigma_b^2)$.

Having a ground truth phenotype, we then compared multiple different genomic prediction models to measure the potential benefits to including the deleterious annotations. Each model followed a similar form:

$$y = X\beta + Z_{block(rep)}b + Z_{rep(trial)}t + Z_A a + Z_D d + \varepsilon$$

This generic model mirrors the previous one, with the exception that germplasm ID is no longer treated as fixed but is instead $\mathbf{Z_A}$ and $\mathbf{Z_D}$are design matrices indicating observations of germplasm IDs for the vectors of additive and dominance effects **a** and **d**, modeled as previously described in the simulated scenario. The six models we compared involve substituting different markers and methods of constructing genomic relationship matrices for $\mathbf{Z_A}$ and $\mathbf{Z_D}$, as well as adding fixed effects for derived homozygous and heterozygous load. The six models include:

- *Marker_G* where the 23,431 variants are used to produce the genomic relationship matrices.
- *Imputed_G* where ~135k imputed genome-wide segregating sites are used to produce the genomic relationship matrices.
- *Nonsyn* where 66k imputed, segregating, nonsynonymous mutation sites are used to produce the genomic relationship matrices.
- *Nonsyn + Load* which is identical to *Nonsyn* with the exception of including the derived load as fixed effects in the prediction
- *Weighted_Nonsyn* uses the same sites as *Nonsyn*, however the genomic relationship matrices are created using the weighted method described previously, with the deleterious weights for each SNP.
- *Weighted_Nonsyn + Load* which is identical to the *Weighted_Nonsyn* with the exception of including the derived load as fixed effects in the prediction

Each model was evaluated by performing the cross-population and within-population predictions as previously described and using the correlation between predicted phenotype and the BLUE as the prediction accuracy (Figures 6, 7). Prediction accuracy was also calculated as the number of the top 25 performing clones predicted as being among the top 25 performing clones (Supplementary Figures S6, S7).

For all simulated scenarios and for empirical within population cross-validations, 95% confidence intervals were calculated. 10-fold cross validation predictions were replicated 30 times, and confidence intervals (CI) were calculated using R:

$$CI = \frac{SD}{sqrt(n)} * qt(p = 0.05/2,\ df = (n-1), lower.tail = F)$$

Where n= # folds * # replications and SD=standard deviation. A true confidence interval assumes observations are independent, which is not true for replications of cross-fold validation, however this gives an estimate for variability in cross-validation prediction accuracies.

4.7 Data availability

Genotype and Phenotype data used in this study is available at cassavabase.org. Euphorbiaceae sequence reads and assemblies generated in this study will be available under bioprojects PRJNA608937 on the Sequence Read Archives and PRJEB55682 on the European Nucleotide Archive, respectively. Code used to process data and produce assemblies, simulations, genomic predictions as well as deleterious weights and mutation results are available at https://bitbucket.org/bucklerlab/cassava_load_and_gp.

Data availability statement

Euphorbiaceae sequence reads and assemblies generated in this study will be available under bioprojects PRJNA608937 on the Sequence Read Archives (SRA) and PRJEB55682 on the European Nucleotide Archive (ENA), respectively.

Author contributions

EL - Collected samples, performed analysis, and did majority of manuscript writing. MR - Managed and organized germplasm collection and genome sequencing. EB - Mentor and oversaw experiments for measuring evolutionary conservation and deleterious mutations, reviewed and edited manuscript. KR - Mentor and oversaw experiments for genomic prediction and deleterious mutations impact on traits, reviewed and edited manuscript. All authors contributed to the article and approved the submitted version.

Funding

This work is supported by workforce development fellowship Project: NYC-149949, Award: 2021-67034-34970 from the USDA National Institute of Food and Agriculture as well as start-up funds from the Robbins lab at Cornell. Additionally, this study is made possible by the funding and

support of the USDA-ARS and the NextGen Cassava project, through the Bill & Melinda Gates Foundation (Grant INV-007637 http://www.gatesfoundation.org) and Commonwealth & Development Office (FCDO).

Acknowledgments

We would like to acknowledge the many germplasm sources that contributed tissue for sequencing (Supplementary Table 1) including: the Denver Botanic Garden, Germplasm Resources Information Network, the Missouri Botanic Garden, the Montgomery Botanic Garden, the National Botanic Garden, the National Tropical Botanic Garden, The New York Botanic Garden, and the US Botanic Garden. Their support was essential in sampling the vast number of species used in this study. We would also like to thank IITA and NaCRRI for contribution of data that we used to cassavabase. In particular, we thank Peter Kulakow, Ismail Rabbi, and Prasad Peteti who were project leads at IITA and Robert Kawuki, a project leader at NaCRRI, and Chiedozie Egesi, overall project manager of the NextGen Cassava Project.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.1041925/full#supplementary-material

References

Agrawal, A. F., and Whitlock, M. C. (2012). Mutation load: The fitness of individuals in populations where deleterious alleles are abundant. *Annu. Rev. Ecol. Evol. Syst.* 43, 115–135. doi: 10.1146/annurev-ecolsys-110411-160257

Alley, E. C., Khimulya, G., Biswas, S., AlQuraishi, M., and Church, G. M. (2019). Unified rational protein engineering with sequence-based deep representation learning. *Nat. Methods* 16 (12), 1315–1322. doi: 10.1038/s41592-019-0598-1

Bachem, C. W. B., van Eck, H. J., and de Vries, M. E. (2019). Understanding genetic load in potato for hybrid diploid breeding. *Mol. Plant* 12, 896–898. doi: 10.1016/J.MOLP.2019.05.015

Bosse, M., Megens, H. J., Derks, M. F. L., de Cara, Á. M. R., and Groenen, M. A. M. (2019). Deleterious alleles in the context of domestication, inbreeding, and selection. *Evol. Appl.* 12, 6. doi: 10.1111/EVA.12691

Browning, B. L., Zhou, Y., and Browning, S. R. (2018). A one-penny imputed genome from next-generation reference panels. *Am. J. Hum. Genet.* 103, 338–348. doi: 10.1016/j.ajhg.2018.07.015

Cheng, H., Concepcion, G. T., Feng, X., Zhang, H., and Li, H. (2021). Haplotype-resolved *de novo* assembly using phased assembly graphs with hifiasm. *Nat. Methods* 18 (2), 170–175. doi: 10.1038/s41592-020-01056-5

Cheruiyot, E. K., Haile-Mariam, M., Cocks, B. G., MacLeod, I. M., Mrode, R., and Pryce, J. E. (2022). Functionally prioritised whole-genome sequence variants improve the accuracy of genomic prediction for heat tolerance. *Genet. Sel. Evol.* 54, 1–18. doi: 10.1186/S12711-022-00708-8/FIGURES/4

de Freitas, J. P. X., da Silva Santos, V., and de Oliveira, E. J. (2016). Inbreeding depression in cassava for productive traits. *Euphytica* 209, 137–145. doi: 10.1007/s10681-016-1649-7

Ferguson, M. E., Shah, T., Kulakow, P., and Ceballos, H. (2019). A global overview of cassava genetic diversity. *PLoS One* 14, 1–16. doi: 10.1371/journal.pone.0224763

Huber, C. D., Kim, B. Y., and Lohmueller, K. E. (2020). Population genetic models of GERP scores suggest pervasive turnover of constrained sites across mammalian evolution. *PloS Genet.* 16, e1008827. doi: 10.1371/JOURNAL.PGEN.1008827

Katoh, K., Misawa, K., Kuma, K. I., and Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* 30, 3059–3066. doi: 10.1093/NAR/GKF436

Khansefid, M., Goddard, M. E., Haile-Mariam, M., Konstantinov, K., Schrooten, C., de Jong, G., et al. (2020). Improving genomic prediction of crossbred and purebred dairy cattle. *Front. Genet.* 11. doi: 10.3389/FGENE.2020.598580

Kono, T. J. Y., Liu, C., Vonderharr, E. E., Koenig, D., Fay, J. C., Smith, K. P., et al. (2019). The fate of deleterious variants in a barley genomic prediction population. *Genetics* 213, 1531–1544. doi: 10.1101/442020

Labroo, M. R., Studer, A. J., and Rutkoski, J. E. (2021). Heterosis and hybrid crop breeding: A multidisciplinary review. *Front. Genet.* 12. doi: 10.3389/FGENE.2021.643761

Li, D., Liu, C.-M., Luo, R., Sadakane, K., and Lam, T.-W. MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly *via* succinct de bruijn graph. *Bioinformatics* 31 (10), 1674–1676. doi: 10.1093/bioinformatics/btv033

Liu, H., Wei, J., Yang, T., Mu, W., Song, B., Yang, T., et al. (2019). Molecular digitization of a botanical garden: high-depth whole-genome sequencing of 689 vascular plant species from the ruili botanical garden. *Gigascience* 8, 1–9. doi: 10.1093/GIGASCIENCE/GIZ007

Loewe, L., and Hill, W. G. (2010). The population genetics of mutations: Good, bad and indifferent. *Philos. Trans. R. Soc. B.: Biol. Sci.* 365, 1153–1167. doi: 10.1098/rstb.2009.0317

Lozano, R., Gazave, E., dos Santos, J. P. R., Stetter, M. G., Valluru, R., Bandillo, N., et al. (2021). Comparative evolutionary genetics of deleterious load in sorghum and maize. *Nat. Plants* 7 (1), 17–24. doi: 10.1038/s41477-020-00834-5

McKey, D., Elias, M., Pujol, M. E., and Duputié, A. (2010). The evolutionary ecology of clonally propagated domesticated plants. *New Phytol.* 186, 318–332. doi: 10.1111/J.1469-8137.2010.03210.X

Moghaddar, N., Khansefid, M., van der Werf, J. H. J., Bolormaa, S., Duijvesteijn, N., Clark, S. A., et al. (2019). Genomic prediction based on selected variants from imputed whole-genome sequence data in Australian sheep populations. *Genet. Sel. Evol.* 51, 72. doi: 10.1186/S12711-019-0514-2

Nasti, R. A., and Voytas, D. F. (2021). Attaining the promise of plant gene editing at scale. *Proc. Natl. Acad. Sci. U.S.A.* 118, e2004846117. doi: 10.1073/PNAS.2004846117/ASSET/F8F17C7C-565B-4915-A746-0A024AC2A114/ASSETS/IMAGES/LARGE/PNAS.2004846117FIG02.JPG

Ng, P. C., and Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.* 31, 3812–3814. doi: 10.1093/nar/gkg509

Nishio, M., and Satoh, M. (2014). Including dominance effects in the genomic BLUP method for genomic evaluation. *PLoS One* 9 (1), e85792. doi: 10.1371/JOURNAL.PONE.0085792

Pan, J. J., and Price, J. S. (2001). Fitness and evolution in clonal plants: The impact of clonal growth. *Evol. Ecol.* 15 (4), 583–600. doi: 10.1023/A:1016065705539

Parmar, A., Sturm, B., and Hensel, O. (2017). Crops that feed the world: Production and improvement of cassava for food, feed, and industrial uses. *Food Secur.* 9, 907–927. doi: 10.1007/s12571-017-0717-8

Ramstein, G. P., and Buckler, E. S. (2022). Prediction of evolutionary constraint by genomic annotations improves prioritization of causal variants in maize. *bioRxiv* 2021, 9.03.458856. doi: 10.1101/2021.09.03.458856

Ramu, P., Esuma, W., Kawuki, R., Rabbi, I. Y., Egesi, C., Bredeson, J., et al. (2017). Cassava haplotype map highlights fixation of deleterious mutations during clonal propagation. *Nat. Genet.* 49, 959–963. doi: 10.1038/ng.3845

Rojas, M. C., Pérez, J. C., Ceballos, H., Baena, D., Morante, N., and Calle, F. (2009). Analysis of inbreeding depression in eight s_1 cassava families. *Crop Sci.* 49, 543–548. doi: 10.2135/cropsci2008.07.0419

Smýkal, P., Nelson, M. N., Berger, J. D., and von Wettberg, E. J. B. (2018). The impact of genetic changes during crop domestication. *Agronomy* 8, 119. doi: 10.3390/AGRONOMY8070119

Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30, 1312. doi: 10.1093/BIOINFORMATICS/BTU033

Valluru, R., Gazave, E. E., Fernandes, S. B., Ferguson, J. N., Lozano, R., Hirannaiah, P., et al. (2019). Deleterious mutation burden and its association with complex traits in sorghum (Sorghum bicolor). *Genetics* 211, 1075–1087. doi: 10.1534/GENETICS.118.301742

van de Peer, Y., Ashman, T. L., Soltis, P. S., and Soltis, D. E. (2021). Polyploidy: An evolutionary and ecological force in stressful times. *Plant Cell* 33, 11–26. doi: 10.1093/PLCELL/KOAA015

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *J. Dairy Sci.* 91, 4414–4423. doi: 10.3168/JDS.2007-0980

Wang, W., Feng, B., Xiao, J., Xia, Z., Zhou, X., Li, P., et al. (2014). Cassava genome from a wild ancestor to cultivated varieties. *Nat. Commun.* 5, 5110. doi: 10.1038/ncomms6110

Wolfe, M. D., del Carpio, D. P., Alabi, O., Ezenwaka, L. C., Ikeogu, U. N., Kayondo, I. S., et al. (2017). Prospects for genomic selection in cassava breeding. *Plant Genome* 10, plantgenome2017.03.0015. doi: 10.3835/plantgenome2017.03.0015

Wright, M. N., and Ziegler, A. (2017). Ranger: A fast implementation of random forests for high dimensional data in c++ and r. *J. Stat. Softw.* 77, 1–17. doi: 10.18637/JSS.V077.I01

Xiang, R., van den Berg, I., MacLeod, I. M., Hayes, B. J., Prowse-Wilkins, C. P., Wang, M., et al. (2019). Quantifying the contribution of sequence variants with regulatory and evolutionary significance to 34 bovine complex traits. *Proc. Natl. Acad. Sci.* 116, 19398–19408. doi: 10.1073/pnas.1904159116

Yang, Z. (2007). PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 24, 1586–1591. doi: 10.1093/MOLBEV/MSM088

Yang, J., Mezmouk, S., Baumgarten, A., Buckler, E., Guill, K., McMullen, M., et al. (2016). Incomplete dominance of deleterious alleles contributes substantially to trait variation and heterosis in maize. *PLOS Genetics* 13 (9), e1007019. doi: 10.1101/086132