Gene Expression Modularity Reveals Footprints of Polygenic Adaptation in *Theobroma cacao*

Tuomas Hämälä,*[,1] Mark J. Guiltinan,[2,3] James H. Marden,[3,4] Siela N. Maximova,[2,3] Claude W. dePamphilis,[3,4] and Peter Tiffin*[,1]

[1]Department of Plant and Microbial Biology, University of Minnesota, St. Paul, MN

[2]Department of Plant Sciences, The Pennsylvania State University, University Park, PA

[3]Huck Institutes of the Life Sciences, The Pennsylvania State University, University Park, PA

[4]Department of Biology, The Pennsylvania State University, University Park, PA

The transcriptome and whole-genome data are available at NCBI SRA: PRJNA558793.

* **Corresponding authors:** E-mails: thamala@umn.edu; ptiffin@umn.edu.

Associate editor: Stephen Wright

Abstract

Separating footprints of adaptation from demography is challenging. When selection has acted on a single locus with major effect, this issue can be alleviated through signatures left by selective sweeps. However, as adaptation is often driven by small allele frequency shifts at many loci, studies focusing on single genes are able to identify only a small portion of genomic variants responsible for adaptation. In face of this challenge, we utilize coexpression information to search for signals of polygenetic adaptation in *Theobroma cacao*, a tropical tree species that is the source of chocolate. Using transcriptomics and a weighted correlation network analysis, we group genes with similar expression patterns into functional modules. We then ask whether modules enriched for specific biological processes exhibit cumulative effects of differential selection in the form of high F_{ST} and d_{XY} between populations. Indeed, modules putatively involved in protein modification, flowering, and water transport show signs of polygenic adaptation even though individual genes that are members of those groups do not bear strong signatures of selection. Modeling of demography, background selection, and the effects of genomic features reveal that these patterns are unlikely to arise by chance. We also find that specific modules are enriched for signals of strong or relaxed purifying selection, with one module bearing signs of adaptive differentiation and an excess of deleterious mutations. Our results provide insight into polygenic adaptation and contribute to understanding of population structure, demographic history, and genome evolution in *T. cacao*.

Key words: **polygenic adaptation, deleterious mutations, background selection, local adaptation, coexpression network, cacao.**

Article

Introduction

The frequency of genomic variants segregating within populations is shaped by both selection and demographic history. Disentangling these effects is complicated given that both adaptation and drift can leave similar signatures at the DNA level (Nielsen 2005). A traditional view arising from population genetics theory is that adaptation is driven by large allele frequency shifts at few loci (Wright 1931), leaving behind distinct signals of "selective sweeps" (Maynard Smith and Haigh 1974). The majority of empirical population genetic analyses have focused on identifying these genes of major effect. By contrast, the contending view of quantitative genetics is that the response to selection happens through correlated changes at many loci—a concept rooted in the "infinitesimal" model of Fisher (1918). The latter process results in subtle signals in population genomic data that are considerably more difficult to detect than sweeps (Stephan 2016). Whether adaptation is attributed to large sweeps, small frequency shifts, or to something in between, is dependent on the genetic architecture underlying phenotypic variation, a species life history, and the strength of selection (Chevin and Hospital 2008; Hermisson and Pennings 2017; Höllinger et al. 2019). Although considerable effort has been spent in quantifying the distribution of effect sizes that individual loci play in human adaptation (Pritchard et al. 2010; Boyle et al. 2017; Sella and Barton 2019), little is known about this topic in other species, particularly in plants.

Regardless of the species, most research on selection shaping genomic diversity has focused on processes that favor the spread of adaptive alleles, that is, positive selection (Biswas and Akey 2006), or to a lesser extent balancing selection (Tian et al. 2002). Recently, however, interest has grown in identifying not only selectively favored alleles but also maladaptive alleles (Chun and Fay 2011; Kono et al. 2016; Zhang et al. 2016; Zhou et al. 2017). These maladaptive alleles constitute a genetic load, and characterizing the extent of the load in a species, as well as the efficacy of purifying selection in removing it, provides insights into the constrains and costs of adaptive evolution (Eyre-Walker and Keightley 2007). In agronomically important species, the identification of

deleterious alleles also may help to understand the costs of domestication and possibly lead to ways for improving yields (Cornejo et al. 2018; Gaut et al. 2018; Valluru et al. 2019).

One approach for resolving the statistical challenges of identifying weak selection acting on multiple genes is to utilize polygenic scores from genome-wide association studies (GWAS) (Berg and Coop 2014; Exposito-Alonso et al. 2018; Rosenberg et al. 2018; Sella and Barton 2019). GWAS, however, requires that adaptive traits are known and measured in large samples. These requirements not only limit the potential application of these methods to a handful of well-studied species but also steer the field toward traits that are relatively easy to measure (e.g., human height: Berg and Coop 2014; Berg et al. 2019; Sohail et al. 2019). A promising alternative that avoids many of the shortcomings of polygenic score analysis is to identify gene groupings, which can be based on putative function (Ashburner et al. 2000), biochemical pathways (Barabási and Oltvai 2004), or coexpression networks (Stuart et al. 2003). Groups that are enriched for signs of nonneutral evolution may thus be responsible for adaptive phenotypes, even if the individual genes that are members of those groups do not carry clear signatures of selection. The task of finding statistical support for the adaptive patterns is still complicated by numerous confounding factors. Besides spurious signals caused by demography (Tiffin and Ross-Ibarra 2014; Hoban et al. 2016), searches for positive selection can be misled by the effects of negative selection on linked variants; a process termed background selection (Charlesworth et al. 1993). Indeed, similar patterns caused by background selection and hitchhiking have led to an effort to include its effects into null models for testing positive selection (Comeron 2017). Controlling for demography and background selection, as well as other features potentially influencing measures of sequence evolution (Cutter and Payseur 2013), is therefore important for realistic modeling of adaptive evolution.

In this study, we analyze genomic data to search for signals of polygenic adaptation in *Theobroma cacao* (hereafter cacao), a tropical tree species best known for being the source of cacao beans. Due to its importance for the chocolate industry, cacao has considerable economic and agronomic value (Guiltinan 2007). Recent work on cacao has provided insight into the species' domestication history (Cornejo et al. 2018; Zarrillo et al. 2018). Here, focusing on wild cacao populations, we first infer the demographic history of this species, and then use transcriptome data to identify cacao-specific coexpression networks. We use these networks to conduct three sets of analyses: First, we search for evidence of specific coexpression modules harboring an excess number of genes that carry strong signatures of adaptive differentiation among populations. Second, we examine the distributions of among-population differentiation estimates to identify modules that may be responsible for local adaptation, even if they do not harbor individual genes bearing strong signals of adaptation. Finally, we search for polygenic signals of relaxed selection by identifying modules showing greater than expected accumulation of deleterious mutations. Our results reveal evidence for both positive and negative selection varying among coexpression modules and provide an example of how polygenic adaptation can be searched for in nonmodel species.

Results

We obtained whole-genome and transcriptome data from 31 cacao individuals, representing four populations: Guiana ($n = 8$), Marañón ($n = 8$), Nanay ($n = 7$), and Iquitos ($n = 8$). These populations originate from diverse locations in central South America (fig. 1*A*) and likely represent the diverse environments in which wild cacao populations are found (Motamayor et al. 2008; Cornejo et al. 2018). Summary statistics estimated from the genome data suggest that these populations have distinct demographic histories (table 1). Based on estimators of population mutation rate ($\theta = 4N_e\mu$), Guiana harbors lower levels of genetic diversity than the other populations, indicating a lower effective population size (N_e). Tajima's *D* estimates were negative in Nanay, possibly due to population size increase after a bottleneck (Tajima 1989), and positive in the other populations, suggesting population size decline. Pairwise F_{ST} estimates and a principal component analysis (PCA) supported earlier findings (Motamayor et al. 2008; Cornejo et al. 2018) by revealing two population groupings: Guiana and Marañón show higher similarity to each other than either does to Nanay or Iquitos, which cluster closely together along the first PC axes (fig. 1*B*). The inferred admixture proportions further support this grouping. Although the likely number of ancestral populations ($K = 4$) corresponds to the four genetic groups, Nanay and Iquitos individuals show higher levels of admixture than individuals belonging to Guiana and Marañón (fig. 1*D*).

To gain further insight into the population history of these four genetic groups, we conducted site frequency spectra (SFS) based demography simulations with fastsimcoal2 (Excoffier et al. 2013). The best supported models (supplementary tables S1 and S2, Supplementary Material online) suggest that the Iquitos population, which is both genetically and geographically close to the Nanay population, split from the other populations as long as 878 K generations ago (fig. 1*C*). Iquitos and Nanay might, however, have been connected by gene flow until ~3.5 K generations ago, which combined with the fairly large and stable effective population sizes (Nanay $N_e \approx 110$ K, Iquitos $N_e \approx 119$ K) may help to explain the relatively low differentiation between them. In contrast, Guiana and Marañón diverged more recently, around 42 K generations ago, but became isolated ~7 K generations ago. They also have lower effective population size estimates (Guiana $N_e \approx 21$ K, Marañón $N_e \approx 63$ K), potentially predisposing them to higher levels of genetic drift. For exact maximum likelihood estimates and their 95% confidence intervals (CI), see supplementary table S3, Supplementary Material online.

Network Enrichment Analysis Reveals the Action of Positive Selection

The four cacao populations we sampled are from areas of South America that differ in temperature, precipitation, and seasonality (supplementary figs. S1 and S2, Supplementary

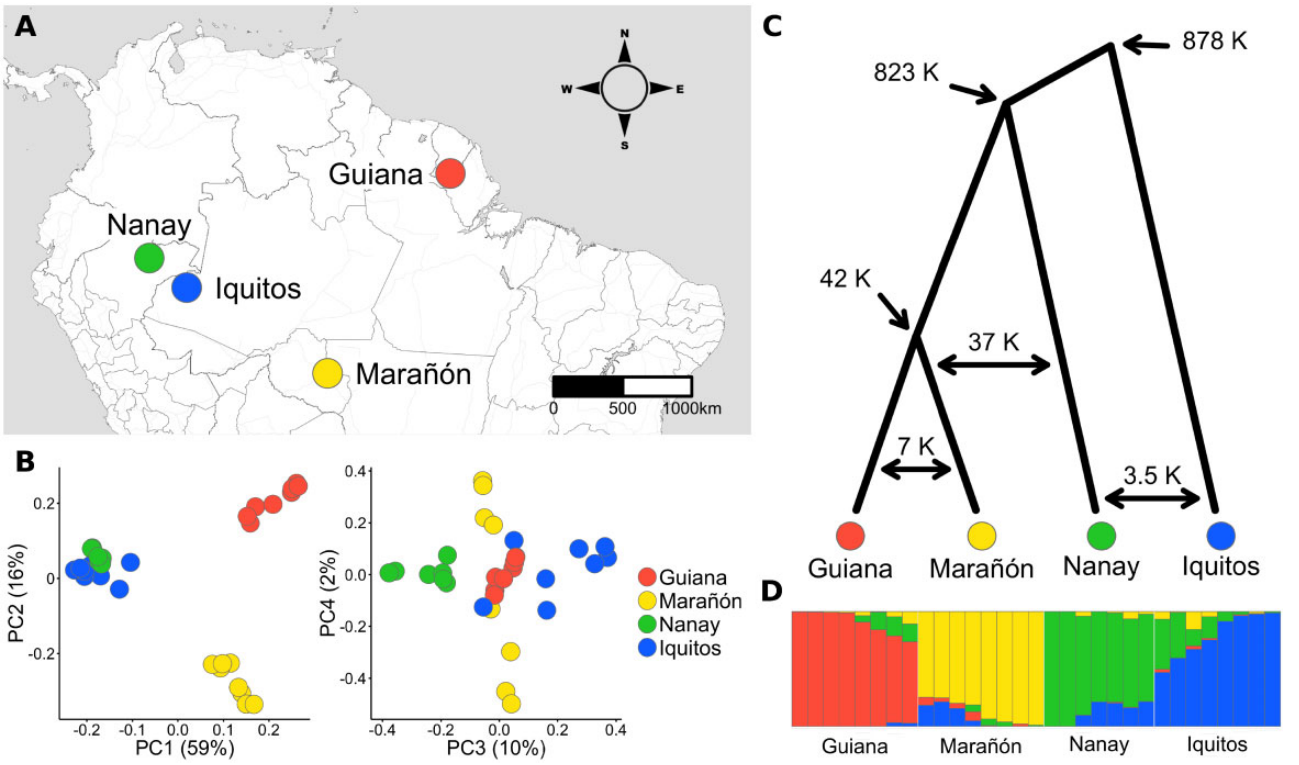

FIG. 1. Population structure and demographic history of the study populations. (*A*) Map showing approximate areas where the populations are found. (*B*) Genetic variation along the first four eigenvectors of a PCA. The variance explained by each PC is shown in brackets. (*C*) A schematic of the demographic history. Divergence times in number of generations, assuming $\mu = 7 \times 10^{-9}$, are marked with one-sided arrows, whereas two-sided arrows indicate times since migration ended. We note that due to uncertainty in mutation rates these estimates might be systematically over- or underestimated. (*D*) Admixture proportions for the supported number of ancestral populations ($K = 4$).

Table 1. Summary Statistics for the Study Populations.

	Genetic Diversity			Pairwise F_{ST}		
	π	θ_w	Tajima's *D*	Guiana	Marañón	Nanay
Guiana	**4.85**	**3.93**	**1.25**			
Marañón	**7.27**	**6.56**	**0.62**	**0.25**		
Nanay	**5.90**	**6.67**	**−0.62**	**0.42**	**0.31**	
Iquitos	**8.82**	**8.47**	**0.25**	**0.36**	**0.27**	**0.18**

NOTE.—π, nucleotide diversity ($\times 10^{-3}$); θ_w, Watterson estimator ($\times 10^{-3}$).

Material online). These populations also differ in their historical effective population sizes and the extent to which gene flow has shaped genomic diversity. To explore whether the action of directional and purifying selection also varies among these populations, as might be expected given that they are found in different geographic areas, we searched for signals of local adaptation at polygenic and single-locus levels. As polygenic selection is characterized by small allele frequency shifts at many loci, the footprints are more readily revealed when considering the cumulative effects of multiple genes underlying adaptive phenotypes. To this end, we utilize an approach previously used with human data (Daub et al. 2013, 2017; Hsieh et al. 2017) by searching for signals of selection among functionally related gene sets. However, cacao gene sets have not been defined on the basis of biochemical pathways or other functional information. We therefore leveraged gene expression variation among our samples to construct a coexpression network to identify groups of genes with correlated expression and then used the coexpression modules as gene sets.

The network we constructed is based on expression variation among genotypes growing in a common environment, rather than variation among tissues or due to the environment in which plants grew. Therefore, the coregulated (or at least correlated) gene expression means that these genes are likely to affect correlated phenotypes, resulting in the selection acting on these phenotypes also being correlated. Because expression patterns do not show strong among-population variation (supplementary fig. S3, Supplementary Material online), we constructed a global coexpression network using data from all individuals. The network consists of 15,073 genes, which are divided into 61 modules (supplementary fig. S4, Supplementary Material online). The number of genes per module ranges from 40 to 1,817, with a median of 122 genes. According to Gene Ontology (GO) enrichment analysis, all modules are enriched for genes involved in specific biological processes (supplementary data set S1, Supplementary Material online), suggesting that the modules capture collections of genes with common functions.

We first focused on identifying modules enriched for individual genes potentially affected by strong directional selection. Out of 20,890 genes, 381 (1.8%) had F_{ST} and d_{XY} estimates that could not be readily explained by the inferred demography (supplementary data set S2, Supplementary

Material online). Two coexpression modules were enriched with these outlier genes ($q < 0.1$, Fisher's exact test). Based on REVIGO (Supek et al. 2011) summarization of the GO terms associated with individual genes (435 and 128 genes, respectively), the largest GO groups were related to defense responses (e.g., response to fungus and wounding) and stress responses (e.g., response to heat and salt stress) (supplementary data set S1, Supplementary Material online). To assess the confidence of these assignments, as well as the confidence of our coexpression network to capture functionally related genes, we also conducted the GO enrichment analysis on permutated data. We examined two features of the randomized GO sets: the number of significant ($q < 0.1$, Fisher's exact test) GO terms and the average P values of these terms. Out of 1,000 randomized repeats, none produced as many significant GO terms as observed in the two modules identified as being enriched for outlier genes. The average P values (reflecting the strength of the GO enrichment) of the random models were, however, lower than was found in these two modules (defense response = 0.00036, stress response = 0.00032, highest P value of the random repeats = 0.00031). This pattern suggests that these two modules contain genes involved in several distinct biological processes. Nevertheless, the fact that genes within these modules exhibited correlated expression indicates that they are coregulated, allowing for selection to act on them jointly.

Neither of the two modules with an excess of outlier genes showed overall deviations from the genome-wide background F_{ST} and d_{XY} (supplementary fig. S5, Supplementary Material online). However, the module putatively involved in stress responses had nucleotide diversity (π), the ratio of nonsynonymous to synonymous diversities (π_N/π_S) and Tajima's D estimates that were lower than expected based on all genes (supplementary fig. S6, Supplementary Material online). Furthermore, for both modules, the ratio of nonsynonymous to synonymous divergence (d_N/d_S) between two cacao subspecies (Motamayor et al. 2013; Argout et al. 2017) was lower than the genome-wide background (supplementary fig. S6, Supplementary Material online). Together, these results suggest that the two modules also experience stronger than average selective constraint.

Small Allele Frequency Shifts at Adaptive Modules

We next evaluated to what degree differential selection among populations drives sequence differentiation of coexpression modules, even if those modules do not contain individual genes bearing signatures of strong selection. To do this, we estimated F_{ST} and d_{XY} for all genes across the four populations, summed the values for genes belonging to each module, and compared the module totals to distribution of values simulated under the best-fit demography model (fig. 1C). We compared the observed values to those from one million simulated gene sets; 3 out of 61 modules exceeded neutral expectations at a nominal q value threshold of 0.01, suggesting that selection has a role in promoting differentiation at these modules. These three modules contained 145, 89, and 130 genes, respectively. It is possible that the empirical observations deviate from the simulated data due to poor fit of the demographic model, but given the very high correspondence between the average observed and simulated estimates (supplementary tables S4 and S5, Supplementary Material online), the model appears to capture neutral patterns reasonably well. However, as the neutral simulations only inform about demography, but not selection, we next compared the observed F_{ST} and d_{XY} estimates to values obtained by randomly assigning genes to modules ($n = 10{,}000$). With these empirical permutations, we retain selection signals at individual genes while breaking up any associations between them. For all three modules, both F_{ST} and d_{XY} estimates exceeded 99% of the permutated repeats ($P < 0.01$, $q < 0.1$; supplementary fig. S7, Supplementary Material online), further supporting that selection on these modules differs from genome-wide expectations. To confirm that these results are not due to a few large-effect genes, we removed individual F_{ST} and d_{XY} outlier genes (17, 4, and 8 genes, respectively) from the modules and compared the adjusted estimates against the genome-wide background. Even after removal of the outlier genes, there remained a low probability ($P < 0.05$, Wilcoxon rank-sum test) of these modules having elevated median values of F_{ST} and d_{XY} by chance. Moreover, for two of the modules, median π, π_N/π_S, and Tajima's D estimates were elevated compared with the genome-wide background (supplementary fig. S6 and data set S3, Supplementary Material online).

A GO enrichment analysis of the three modules with elevated median differentiation revealed that they are enriched for genes involved in protein modification (e.g., protein ubiquitination and cellular protein modification), flowering (e.g., flower development and photoperiodism), and water transport (e.g., water transport and response to water deprivation), respectively (fig. 2A and B and supplementary data set S1, Supplementary Material online). Indeed, in the first module 82 of 92 genes with defined GO terms were involved in protein modification, in the second module 37 out of 53 were involved in flowering, and in the third module 55 out of 71 genes were involved in water transport. These modules were also enriched for fewer GO terms than expected by chance; 88%, 91%, and 98% of permutated repeats ($n = 1{,}000$) produced more significant GO terms than the observed data. By contrast, the statistical strength of the enrichment was stronger than expected; only 10%, 3%, and 1% of the permutations had average P values that were lower than the observed P values.

Because genomic features affect sequence evolution (Begun and Aquadro 1992; Nordborg et al. 1996; Cutter and Payseur 2013; Mähler et al. 2017; Hämälä and Savolainen 2019), it is possible that the signals of elevated differentiation are biased by the genomic neighborhood in which the genes are found. To explore this possibility, we tested whether genes at the candidate modules are atypical in terms of recombination rate, gene density, mutation rate, connectivity, expression level, expression variance, and B (a proxy for the strength of background selection, Hudson and Kaplan 1995; Nordborg et al. 1996). Although both F_{ST} and d_{XY} were correlated with the strength of background selection (B), recombination rate, and gene density (table 2),

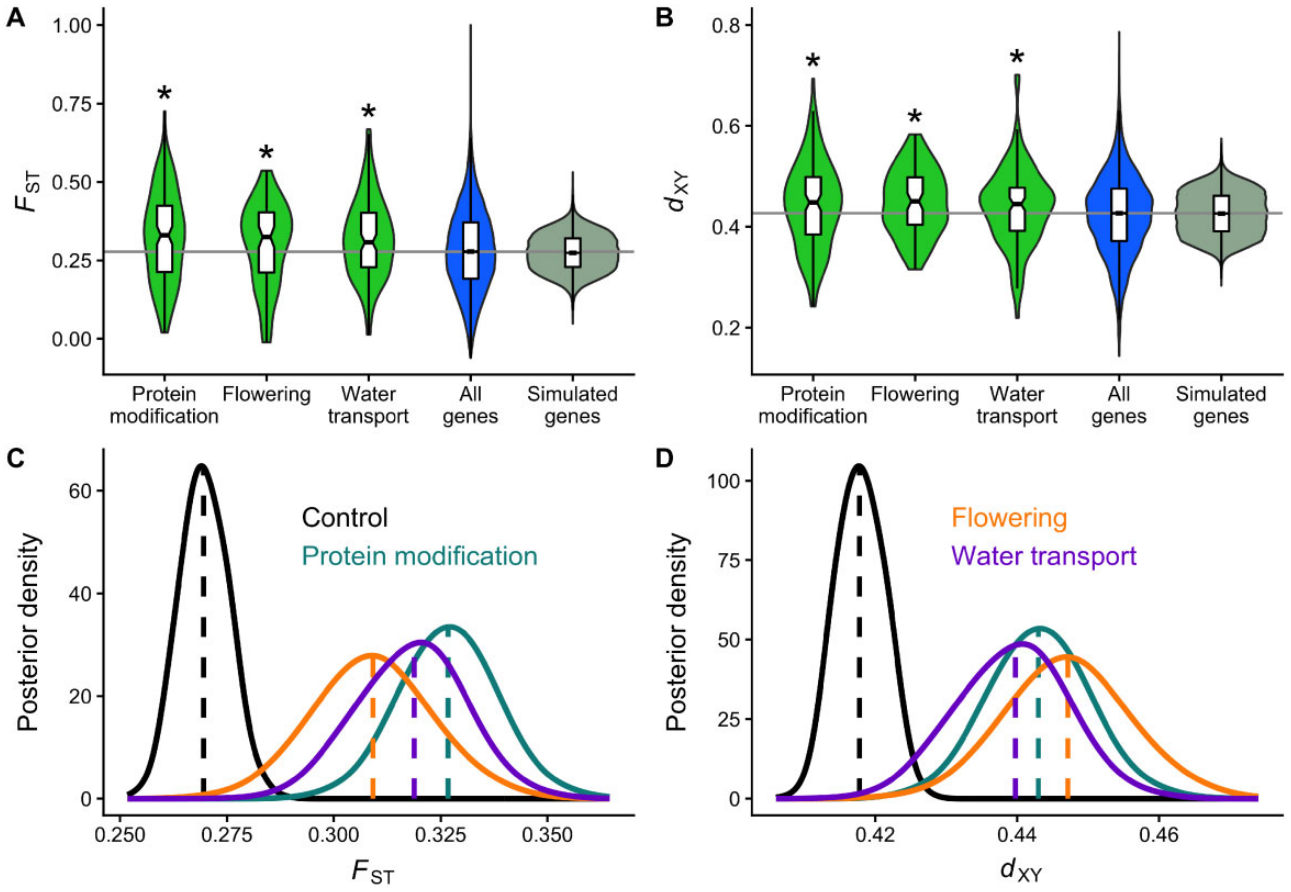

FIG. 2. Measures of differential selection at candidate modules. (*A* and *B*) The sampling distributions of F_{ST} and d_{XY} compared with observed genome-wide distribution and the full simulated distribution. The gray horizontal line marks the median of the genome-wide distribution and stars indicate significant differences in comparison to that ($P < 0.05$, Wilcoxon rank-sum test). (*C* and *D*) Posterior density distributions from Bayesian linear models which control for genomic features. The candidate modules are compared against 500 randomly sampled genes (control, also note color codes apply to *C* and *D*).

distributions for the candidate modules did not clearly differ from the genome-wide background for any of them ($P > 0.1$, Wilcoxon rank-sum test). Bayesian linear regression modeling also revealed that genetic differentiation was affected by *B*, recombination rate, and gene density (note that even though these variables are correlated at the genome-wide level [supplementary table S6, Supplementary Material online], models showed no collinearity between them, $\sqrt{\text{variance inflation factor}} < 2$). Nevertheless, when compared against the genome-wide background, regression coefficients of the three candidate modules remained significant (95% highest posterior density [HPD] intervals >0) in a model with *B*, recombination rate, and gene density included (fig. 2*C* and *D*). Furthermore, out of 10,000 repeats where genes were randomized across modules, <1% produced HPD intervals as extreme as the observed models. All of these results suggest that these modules experience polygenic selection that differs from genome-wide expectations.

Accumulation of Deleterious Mutations at the Module Level

We next explored whether modules vary in their accumulation of deleterious mutations, reflecting maladaptive hitchhiking or relaxed selective constraint. To this end, we utilized patterns of protein conservation among homologous sequences to predict mutational effects with SIFT4G (Vaser et al. 2016). Genome-wide, there appears to be a large accumulation of putatively deleterious mutations within cacao, as the SIFT-based prediction of the ratio of deleterious to synonymous variants was 0.35. SIFT scores are not, however, random with respect to coexpression modules. As with F_{ST} and d_{XY}, gene specific estimates from SIFT were summed for each module and compared against randomly compiled modules of equal size ($n = 10{,}000$). Of the 61 modules, 3 were enriched for low SIFT scores among all individuals, suggesting that these modules harbor an excess of deleterious mutations (fig. 3*A*). SIFT scores estimated for each population confirmed that the harmful mutations at these modules are not specific to just one of the populations, as two of these modules had a significant excess ($q < 0.1$, Fisher's exact test) of low SIFT scores in all four populations and one module was significantly enriched in three of the populations (Iquitos, Guiana, and Marañón). Regression modeling revealed that SIFT scores are negatively related to recombination rate and expression variance, and positively related to mutation rate and expression level (table 2). However, these features are unlikely responsible for the excess of deleterious mutations at

Table 2. The Relationship between Genomic Features and Measures of Sequence Evolution.

	F_{ST}		d_{XY}		SIFT		π		π_N/π_S		Tajima's *D*	
	ρ	β	ρ	β	ρ	β	ρ	β	ρ	β	ρ	β
Background selection	−0.15*	−0.07*	−0.10*	−0.05*	−0.01	0.01	0.24*	0.20*	0.01	0	0.05*	0.04*
Recombination rate	−0.13*	−0.10*	−0.05*	−0.04*	−0.05*	−0.04*	0.25*	0.13*	0.05*	0.05*	0.11*	0.05*
Gene density	−0.03*	−0.05*	−0.05*	−0.06*	0.02	0.02	−0.06*	−0.01	−0.01	−0.01	−0.02*	−0.02*
Mutation rate	−0.01	−0.01	−0.01	0.02	0.11*	0.06*	0.17*	0.09*	−0.14*	−0.05*	0.05*	0.05*
Connectivity	0.01	0	−0.01	−0.02	0.02	0	−0.09*	−0.08*	0.01	0	−0.05*	−0.04*
Expression level	0.01	0.01	−0.01	−0.01	0.16*	0.16	−0.10*	−0.08*	−0.13*	−0.12*	−0.05*	−0.04*
Expression variance	0	−0.01	0.01	0.01	−0.09*	−0.06*	0.13*	0.11*	0.10*	0.06*	0.06*	0.05*

NOTE.—For full correlation matrix, see supplementary table S6, Supplementary Material online. ρ, Spearman's rank correlation coefficient; β, coefficients from multiple regression models.
*$P < 0.05$ (ρ) or BF>1 (β).

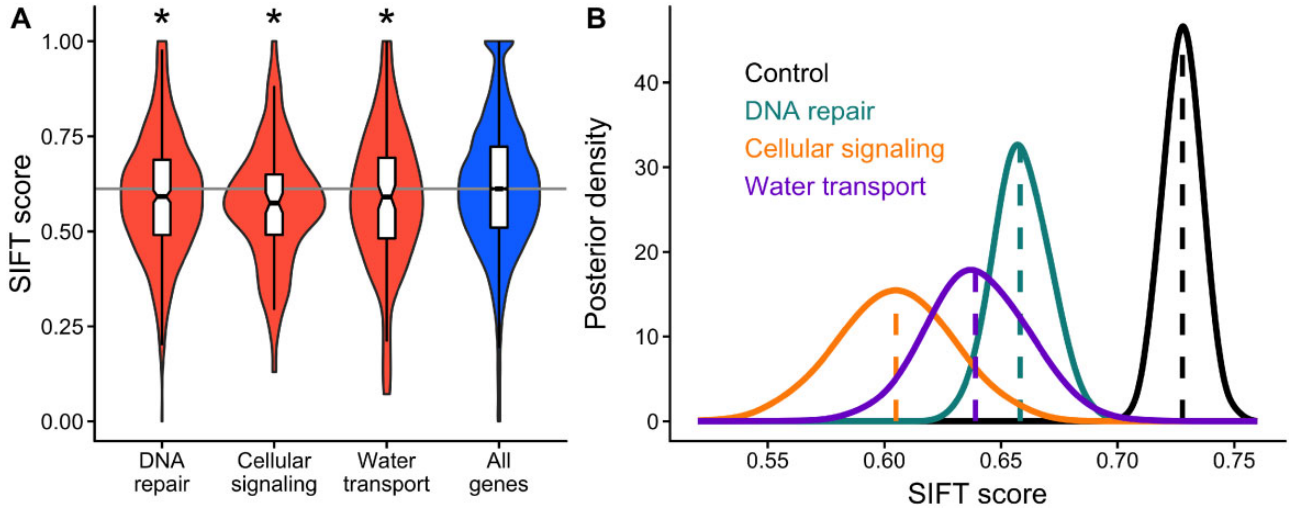

FIG. 3. SIFT scores at candidate modules. (*A*) The module-distributions compared with the genome-wide distribution. The gray horizontal line marks the median of the genome-wide distribution and stars indicate significant differences in comparison to that ($P < 0.05$, Wilcoxon rank-sum test). (*B*) Posterior density distributions from a Bayesian linear model that controls for genomic features. The candidate modules are compared against 500 randomly sampled genes.

these modules (fig. 3*B*). Among the three modules harboring an excess of deleterious mutations was the water transport module that also was characterized by elevated estimates of F_{ST} and d_{XY}. The other two SIFT outlier modules were enriched for genes involved in DNA repair (100 genes) and cellular signaling (438 genes) (supplementary data set S1, Supplementary Material online). However, 74% and 19% of the permutations ($n = 1{,}000$) produced more significant GO terms, and 14% and 91% had lower average *P* values than the observed data, suggesting these two modules might be involved in multiple biological processes.

To seek additional validation for the inferred accumulation of harmful mutations in specific modules, we estimated the distribution of fitness effects (DFE) with DFE-alpha (Keightley and Eyre-Walker 2007). As opposed to SIFT predictions, DFE is based on comparing the SFS of nonsynonymous (0-fold) to synonymous (4-fold) sites. DFE estimated from the whole-genome SFS was strongly L-shaped (shape parameter of the gamma distribution = 0.06), meaning that when nonsynonymous sites are assigned to bins based on the strength of purifying selection ($N_e s$), a large proportion of variants have either nearly neutral ($N_e s < 1$) or highly deleterious ($N_e s > 100$) estimates (fig. 4). Population-specific DFE differed slightly from one another, consistent with the inferred demography (e.g., Guiana, the population with the smallest N_e, has a greater proportion of sites in the nearly neutral category), but there were no major differences among them (supplementary fig. S8, Supplementary Material online). Among the modules identified in our previous analyses, the two modules enriched among F_{ST} and d_{XY} outlier genes (defense response and stress response) were enriched for sites in the highly deleterious category ($N_e s > 100$). These modules also had lower than expected d_N/d_S estimates (supplementary fig. S6, Supplementary Material online), indicative of stronger than average selective constraint (fig. 4). Two of the three modules with high median F_{ST} and d_{XY} estimates (protein modification and flowering) had DFE similar to the whole-genome background (fig. 4). The third outlier module (water transport), as well as the other two modules with low SIFT scores, bore multiple signatures of relaxed selective constraint. In particular, each of these three modules had an excess of sites in the nearly neutral category ($N_e s < 1$), as well as π_N/π_S

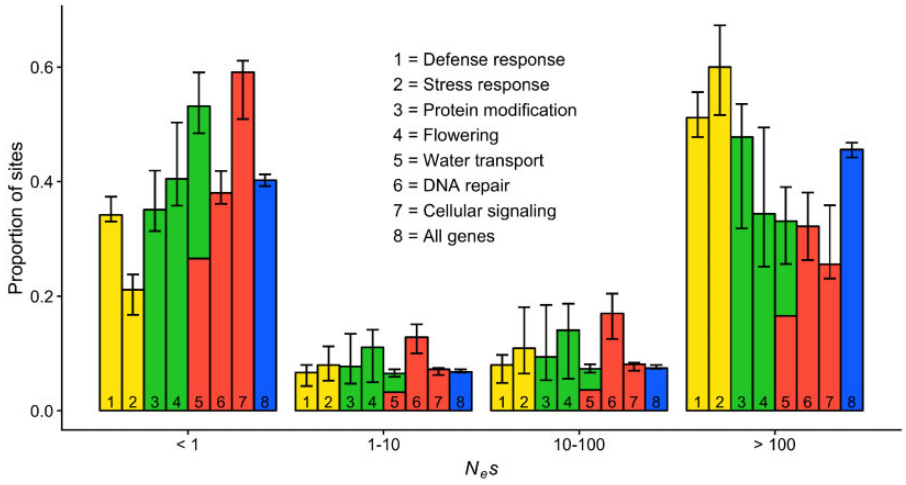

FIG. 4. The DFE for candidate modules. The 0-fold degenerate sites are divided into four bins based on the strength of purifying selection ($N_e s$). Error bars show 95% CIs. The modules of interest were detected using different methods: 1 and 2, enriched among F_{ST} and d_{XY} outlier genes; 3–5, enriched for high F_{ST} and d_{XY}; 5–7, enriched for low SIFT scores; 8, all 15,073 genes in the transcriptome data.

and d_N/d_S estimates that were higher than the genome-wide background (supplementary fig. S6, Supplementary Material online).

Discussion

Genomic Footprints of Polygenic Adaptation

The growing recognition that adaptation is rarely attributable to only few large-effect genes, along with advances in sequencing technologies, has led to an increase in studies examining genomic signatures of polygenic adaptation (Pritchard et al. 2010; Daub et al. 2013, 2017; Berg and Coop 2014; Boyle et al. 2017; Hsieh et al. 2017; Sella and Barton 2019). Most of this work has focused on humans, although a few studies have been done with plant species (He et al. 2016; Beissinger et al. 2018; Exposito-Alonso et al. 2018; Josephs et al. 2019). A common way to search for the quantitative signals of selection is by obtaining polygenic scores from GWAS (Rosenberg et al. 2018). Although this method can have high sensitivity when applied to large samples, the focus on preselected phenotypes might lead to exclusion of important traits. Furthermore, without controlling for the environmental differences, polygenic scores are easily biased by population structure (Berg et al. 2019; Sohail et al. 2019). An alternative to polygenic scores is to examine the properties of genes with shared function (Daub et al. 2013, 2017; Hsieh et al. 2017). Here, we used transcriptome data to identify groups of genes with coregulated expression and then used the coexpression modules to search for genomic footprints of polygenic adaptation among cacao populations.

Our analyses identified two sets of modules with potential signatures of polygenic adaptation. One of these sets was comprised of two modules that contained an excess of genes with signals of strong selection. GO enrichment analysis revealed that these modules are most strongly enriched for defense and stress response genes. However, the analyses also revealed enrichment for more biological processes than expected by chance, suggesting that the modules may actually capture genes involved in several distinct physiological or developmental processes. Nevertheless, the among-genotype correlations in gene expression suggest that variation in the phenotypes associated with each of these modules is at least partially correlated. Regardless of the phenotypic traits associated with these modules, the excess of highly differentiated (F_{ST} and d_{XY}) genes corresponds to a model where adaptive differentiation advances though large, but still polygenic, allele frequency shifts (Höllinger et al. 2019). Interestingly, these two modules also show evidence (SIFT scores, DFE, and d_N/d_S) of strong selective constraint. This combination of results suggests a model in which the majority of genes in these modules are under selective constraint, leaving few that are unconstrained and thus able to respond to selection.

The other set of modules bearing a potential signature of polygenic adaptation was comprised of three modules with F_{ST} and d_{XY} distributions shifted toward large estimates. Importantly, genes targeted by strong selection were not overrepresented among these modules, but the footprint of selection became apparent only when considering the cumulative effects of multiple genes, each bearing only subtle signals. This pattern is, in fact, the hallmark of polygenic adaptation (Chevin and Hospital 2008; Stephan 2016; Höllinger et al. 2019). Based on GO enrichment analysis, these modules are potentially involved in protein modification, flowering, and water transport. Protein modification is not among the classical traits related to local adaptation, but it undoubtedly has a role in myriad processes, making it a viable target for differential selection (Collevatti et al. 2019). Signatures of adaptation at modules related to both flowering and water transport are consistent with previous work on local adaptation (Barrett and Hoekstra 2011; Savolainen et al. 2013; Tiffin and Ross-Ibarra 2014). Moreover, the cacao populations are from areas that differ in temperature and precipitation (supplementary figs. S1 and S2, Supplementary Material online), both of which may impose selection that

affects flowering time and traits associated with water availability. Although genetic studies have discovered major effect loci underlying flowering time variation in some plant species (Salomé et al. 2011; Keller et al. 2012; Wang et al. 2018), common-garden experiments, and GWAS on other species point to variation being due to many genes with small effects (Savolainen et al. 2007; Buckler et al. 2009; Hämälä et al. 2018). Similarly, the genetic basis of variation in response to water limitation, which is likely shaped in part by physiological processes related to water transport, are often highly polygenic (He et al. 2016; Exposito-Alonso et al. 2018).

Polygenic Adaptation and Accumulation of Deleterious Mutations

To examine the role of purifying selection in cacao adaptation, we used two complementary approaches for inferring deleterious effects: sequence homology-based prediction with SIFT4G (Vaser et al. 2016) and estimation of fitness effects with DFE-alpha (Keightley and Eyre-Walker 2007). At the whole-genome level, cacao harbors a high proportion of putatively deleterious alleles (based on SIFT predictions), a high ratio of nonsynonymous to synonymous nucleotide diversities ($\pi_N/\pi_S = 0.36$), and a L-shaped DFE—all of which are indicative of relaxed purifying selection (Eyre-Walker and Keightley 2007; Chen et al. 2017). Such patterns are commonly observed in species that have undergone a domestication bottleneck (Kono et al. 2016; Gaut et al. 2018; Valluru et al. 2019), but they also can result from other processes. For example, comparisons of DFE across multiple species have found that greater longevity and lower outcrossing rate generally result in increased proportion of nearly neutral mutations (Gossmann et al. 2010; Chen et al. 2017). Cacao, being a long-lived insect-pollinated tree species that has been utilized by humans for thousands of years (Zarrillo et al. 2018), might therefore exhibit relaxed purifying selection due to both natural and anthropogenic reasons.

We then examined how harmful mutations are distributed among the coexpression modules to evaluate whether some modules harbor an enrichment of deleterious mutations, and if there is a relationship between the efficacy of purifying selection and polygenic adaptation. The traditional model of hitchhiking states that probability of highly deleterious mutations being fixed is correlated with the strength of positive selection acting on linked variants (Maynard Smith and Haigh 1974). Evidence for hitchhiking of deleterious variants has been found in several domesticated plant species (Kono et al. 2016; Gaut et al. 2018; Valluru et al. 2019), and hitchhiking appears to be responsible for increased deleterious load in humans (Chun and Fay 2011) and in poplar (Zhang et al. 2016). In contrast to most previous studies, here, we did not focus only on highly deleterious mutations (e.g., SIFT score <0.05). By looking at all putatively harmful mutations, our analyses have the potential to reveal evidence for the accumulation of deleterious alleles that might result from hitchhiking associated with polygenic adaptation or processes that result in relaxed purifying selection.

We found that five modules were outliers in terms of overall SIFT scores; two modules with high SIFT scores, indicative of stronger purifying selection and three modules with low SIFT scores, indicative of relaxed selection. The two modules bearing signatures of high selective constraint also harbored an excess number of genes with high F_{ST} and d_{XY} estimates. This combination suggests that most genes in these modules (defense and stress response) are highly constrained, but that there are evolutionary labile parts of the module that are important for adaptation. Interestingly, one of the three modules with a low SIFT scores, indicative of an excess of deleterious mutations, is also characterized by elevated median F_{ST} and d_{XY} estimates. This module, putatively involved in water transport, also exhibited an elevated proportion of nearly neutral mutations, and greater than expected π_N/π_S and d_N/d_S estimates.

At first glance it seems that this pattern might fit the model of maladaptive hitchhiking, in which positive selection increases the frequency of linked deleterious variants. However, as both theoretical (Maynard Smith and Haigh 1974; Barton 1995; Hartfield and Otto 2011) and empirical (Chun and Fay 2011) studies have associated maladaptive hitchhiking with high selection coefficients, weak selection acting on multiple loci seems unlikely to be able to increase the deleterious load. Moreover, genetic hitchhiking requires physical linkage between positively selected and deleterious variants, whereas genes that comprise the coexpression modules are not physically linked in the genome. As such, if genetic hitchhiking is responsible for the excess of deleterious mutations, we would expect a negative relationship between the measures of positive selection (F_{ST} and d_{XY}) and SIFT scores. However, these measures were effectively uncorrelated (Spearman's $\rho \approx 0$) in our data. The lack of correlation suggests that the signatures of relaxed purifying selection at this module are due to a recent change in selective constraint (Chun and Fay 2011). Such change could, for example, happen if the selection pressure on traits controlled by these loci has lessened due to environmental change. The increased genetic drift that follows the decrease of purifying selection may then allow random alleles to reach high frequencies, resulting in increased F_{ST} and d_{XY} estimates. Although our analysis shows that these high estimates are unlikely to arise by chance, it is important to realize that these tests are made against the hypothesis that selection is not acting differently on any group of genes. In other words, if genes from a single module are exposed to high drift due to relaxed selective constraint, it might lead to apparent signals of positive selection that exceeds genome-wide expectations.

The other two modules showing evidence of relaxed purifying selection, one possibly involved in DNA repair and one possibly involved in cellular signaling, did not show any signs of having been subject to positive selection. Therefore, the increase in deleterious load at these modules is also likely the result of relaxed selective constraint (Chun and Fay 2011). Previous work has found increased deleterious load in the domesticated cacao cultivar, with evidence of decreasing yield due to this load (Cornejo et al. 2018). Here, we have shown that compared with many other undomesticated plant species (Gossmann et al. 2010; Chen et al. 2017), the wild cacao populations also exhibit relaxed purifying selection, and that

the deleterious mutations are nonrandomly distributed among the coexpression modules, which is likely related to relaxed selective constraint.

Evidence of Linked Selection across the Cacao Genome

Our analysis of genomic features not only revealed potential factors confounding searches for selection but also provided insight into genome evolution in cacao. Consistent with the evidence of relaxed purifying selection, the mean estimate of $B = 0.90$ suggests that background selection at genic regions is not a major factor underlying diversity patterns in cacao. However, the range of B estimates ($0.1 < B < 1.0$) indicates that the effect on some genes can be considerable, and that all genes are affected by background selection (supplementary fig. S9, Supplementary Material online). Indeed, our analysis showed that measures of sequence differentiation are influenced by background selection, with F_{ST} being more strongly correlated with B than d_{XY}. As the former is a relative measure of differentiation ($F_{ST} = 1-H_W/H_B$, where H_W and H_B are the mean number of differences within and between population, Hudson et al. 1992), it can be biased when background selection reduces diversity without the accompanying increase in divergence (Charlesworth 1998; Cruickshank and Hahn 2014). Our results therefore support the conclusion that controlling for background selection is beneficial in searches for positive selection (Comeron 2017).

To have a more complete picture of factors driving genome evolution in cacao, we also examined how individual genomic features affect genetic diversity. The regression modeling of π revealed that diversity is positively associated with recombination rate, expression variance, and mutation rate, but negatively associated with expression level, connectivity, and gene density. These relationships mostly stem from the effects of positive and negative selection on linked variants (Maynard Smith and Haigh 1974; Begun and Aquadro 1992; Charlesworth et al. 1993; Hudson and Kaplan 1995; Nordborg et al. 1996; Cutter and Payseur 2013). Namely, low recombination increases the extent of linked selection, and high gene density creates a larger target for selection to act on (Hudson and Kaplan 1995; Nordborg et al. 1996; Cutter and Payseur 2013). Recombination might, however, be mutagenic itself, so controlling for mutation rate variation through neutral divergence is important for disentangling these effects (Begun and Aquadro 1992; Cutter and Payseur 2013). On the other hand, high connectivity and expression level are usually positively associated with stronger selective constraint (Josephs et al. 2017; Mähler et al. 2017).

Conclusions

Here, we have presented a novel approach for finding footprints of polygenic adaptation that is not dependent on preselected phenotypes or predefined gene sets. In fact, the only requirement is transcriptome data from an appropriate population sample, making the method suitable for both model and nonmodel species. By applying measures of sequence evolution on gene sets derived from a coexpression network, we have gained insights into polygenic adaptation in a tropical tree species, cacao. In particular, we found that evidence for both adaptive differentiation among cacao populations, and the effects of purifying selection are nonrandomly distributed among coexpression modules. Three modules, those putatively involved in protein modification, flowering, and water transport, showed orchestrated allele frequency shifts consistent with polygenic selection that may have accompanied climate change or range expansion into more seasonally variable climates in central and eastern Amazonia. One of these modules, involved with water transport, also bears an excess of deleterious mutations, suggesting relaxed purifying selection that became evident when examining genes at the level of coexpression modules. The increasing availability of gene expression and genomic data for diverse species, combined with the generally polygenic nature of adaptation and the difficulty of disentangling signals of selection from demographic effects, makes the approach presented here a promising way to quickly identify gene sets likely to be involved in local adaptation.

Materials and Methods

Sample Collection, Library Preparation, and Sequencing

We acquired whole-genome and transcriptome data from 31 cacao individuals, which represented four populations: Guiana ($n = 8$), Marañón ($n = 8$), Nanay ($n = 7$), and Iquitos ($n = 8$). Leaf tissue was collected from trees maintained in a common garden in the International Cacao Collection at Centro Agronómico Tropical de Investigación y Enseñanza (CATIE), Costa Rica. Two healthy and mature (stage E) leaves were collected from each genotype, flash frozen, and ground under liquid nitrogen. DNA was extracted by taking ground tissue powder at low input (0.25 g/15 ml lysis buffer) through a CTAB isolation based on Michiels et al. (2003) with the following modifications: the lysis buffer contained 4% PVP; the isopropanol precipitation was carried out at −20°C overnight; the day 2 clean up steps were carried out in 50 ml tubes; and proteinase K digest (20 μl/ml Qiagen #19131, 2 h at 50°C) was added prior to the phenol: chloroform: isoamyl alcohol extraction. Final pellets were resuspended in 150 μl TE. RNA was extracted from the ground tissue according to the PureLink Plant RNA Reagent protocol (Thermo Fisher Scientific), with minor modifications as previously described in Pokou et al. (2019). After extraction, RNA samples were treated with RNase-free DNase (Thermo Fisher Scientific) according to manufacturer's instructions, purified with an additional ethanol precipitation, and resuspended in RNase-free water. Library preparation and sequencing were performed at The Pennsylvania State University Genomics Core Facility. Illumina TruSeq DNA PCR-Free High Throughput kit and Illumina TruSeq Stranded mRNA Library kit were used to prepare the DNA and RNA libraries, respectively. Pooled libraries were normalized and denatured with 0.2 N NaOH for a loading concentration of 375 pM on a NovaSeq S2 (300) flowcell. Cluster amplification of denatured

templates and 150-bp paired-end sequencing was performed according to the standard Illumina NovaSeq S2 protocol.

Processing of the Whole-Genome Sequencing Reads

Low-quality reads and sequencing adapters were removed with Trimmomatic (Bolger et al. 2014), and the remaining reads were aligned to the cacao reference genome, Criollo v2.0 (Argout et al. 2017), with BWA-MEM (Li and Durbin 2009; Li 2013). GATK (DePristo et al. 2011) was used to remove duplicated reads, add read group information and realign indels. The median read coverage per genome per individual ranged from 21 to 46×. Even with these high medians, the variability in coverage might introduce bias into our results (Nielsen et al. 2011). We therefore took a genotype call-free approach with most of the downstream analyses. Genotype likelihoods were generated with the GATK model in ANGSD (Korneliussen et al. 2014) to be used instead of SNP-calls. However, to estimate population recombination rates (see "Detection of Differential Selection"), genotype calling was required. To this end, variants called with Freebayes (Garrison and Marth 2012) were filtered with VCFtools (Danecek et al. 2011) and phased with Beagle 5.0 (Browning et al. 2018). With both genotype likelihoods and genotype calls, sites were required to have a mapping quality over 30 and site quality over 20. SNP-calls with genotype quality <20 and coverage <10× were further removed.

Analysis of Population Structure

Genetic diversity within populations was estimated using three summary statistics with ANGSD (Korneliussen et al. 2014): nucleotide diversity π (Tajima 1983) and Watterson estimator θ_W (Watterson 1975), which are estimators of the population mutation rate $\theta = 4N_e\mu$, and Tajima's *D*, which can indicate skews in the site frequency spectrum due to demography or selection (Tajima 1989). We then evaluated genetic relationships among the study populations by conducting a PCA and an admixture analysis with PCAngsd (Meisner and Albrechtsen 2018). The supported number of PCs was used to define the likely number of ancestral populations (*K*).

Demography Simulations

To study the demographic history of these populations, we conducted SFS-based coalescent simulations with fastsimcoal2 (Excoffier et al. 2013). Folded SFS were estimated from 4-fold degenerate sites with ANGSD (Korneliussen et al. 2014). As the multidimensional SFS could be reliably estimated only for three populations, we performed demography simulations using all possible 3 D combinations. Simulations with different divergence orders and migration scenarios were repeated 50 times and the relative fit among models with highest likelihoods was tested with the Akaike information criterion (AIC). The best fitting models were used to define effective populations sizes (N_e), divergence times (*T*), and migration rates (*m*) for the populations. CIs for the parameters were estimated using 100 nonparametric bootstrap SFS. Mutation rate $\mu = 7 \times 10^{-9}$ per base pair was assumed for all simulations. This estimate is derived from mutation accumulation experiments in *Arabidopsis thaliana* (Ossowski et al. 2010) and it is close to one estimated with parent–offspring sequencing for a woody perennial, peach (7.77×10^{-9}, Xie et al. 2016). The estimate is not, however, cacao-specific and the uncertainty in mutation rates affects our demography estimates at the overall level (e.g., higher μ would lead to lower N_e and *T* estimates), but it does not influence the relative differences between populations.

Construction of a Coexpression Network

After removing low-quality reads and sequencing adapters with Trimmomatic (Bolger et al. 2014), STAR (Dobin et al. 2013) was used to align RNA-Seq reads against the Criollo v2.0 reference genome, and count reads overlapping each gene model. We only used genes with median read count over 10. Count data were normalized with variance stabilization transformation (VST) in DESeq2 (Love et al. 2014) prior to expression analysis. The presence of population structure in the expression data was evaluated with PCA in base R (R Core Team 2019). We then used weighted correlation network analysis (WGCNA) (Langfelder and Horvath 2008) to detect genes with similar expression patterns among samples. Based on the criterion of an approximate scale-free topology, a soft thresholding power of 12 was used to calculate adjacencies for a signed coexpression network. Topological overlap matrix (TOM) and dynamic-cut tree algorithm were used to detect network modules. We imposed a minimum module size of 30 genes for the initial network construction and later merged modules exhibiting >90% similarity. To infer putative functions for these modules, we performed BLAST queries against the *A. thaliana* nucleotide database (Berardini et al. 2015). The GO terms (Ashburner et al. 2000) associated with *A. thaliana* homologs (alignment *e* value $<1 \times 10^{-5}$) were extracted and used to find common biological processes among cacao genes belonging to the same modules. One-sided Fisher's exact test was used to detect overrepresented terms, which were summarized with REVIGO (Supek et al. 2011). We refer to modules by their largest REVIGO group, although we acknowledge that many genes in these modules may be involved in other processes.

Detection of Differential Selection

We estimated allele frequency differentiation between populations to identify genes potentially contributing to local adaptation. After employing a model by Kim et al. (2011) to estimate allele frequencies from the genotype likelihoods, differentiation proportions were calculated with a relative measure F_{ST} (Wright 1951) and an absolute measure d_{XY} (Nei 1987). In contrast to F_{ST}, d_{XY} is not affected by within-population levels of genetic diversity (Charlesworth 1998; Cruickshank and Hahn 2014), but it can be biased by unequal sample sizes. We therefore focused on genes and modules showing signs of selection with both measures. We used information from SNPs that localized within the Criollo v2.0 gene models to define gene specific F_{ST} and d_{XY} estimates. F_{ST} was estimated with Hudson's measure (Hudson et al. 1992; Bhatia et al. 2013), using a weighting method by Reynolds

et al. (1983) to combine estimates across multiple sites. d_{XY} was estimated for SNPs, as opposed to all sites within a gene, so as not to bias the estimates based on different combinations of variable sites and gene lengths. Per gene F_{ST} and d_{XY} were calculated for each population comparison and averaged to acquire global estimates. To identify genes showing higher than neutral differentiation, the observed values were compared with F_{ST} and d_{XY} values from simulated data generated with *ms* (Hudson 2002). To obtain an approximation of recombination rate variation across the Criollo v2.0 genome, we used the genomic data and FastEPRR (Gao et al. 2016) to estimate population recombination rates ($4N_er$) in 50-kb sliding windows. We performed this analysis for each population separately, used the N_e parameters to transform $4N_er$ to r, and averaged the per base pair recombination rates across the four populations. We used these r estimates in combination with N_e, m, and T, going back to most recent common ancestor of these populations, to generate 200,000 neutral fragments ($\sim$10$\times$ the number of protein coding genes in the transcriptome data) that corresponded in size to gene lengths obtained from the genome annotation. We defined P values by comparing the observed estimates against percentiles of the simulated distributions. The P values were adjusted for multiple testing by transforming them into false discovery rate-based q values (Storey and Tibshirani 2003). Genes with F_{ST}- and d_{XY}-based q values $<$0.01 were considered putatively adaptive.

Analysis of Deleterious Mutations

Putatively deleterious mutations were identified with SIFT4G (Vaser et al. 2016). We built a custom database for the Criollo v2.0 reference genome by comparing the translated protein sequences against UniRef90 database (Suzek et al. 2015). Based on protein conservation among homologous sequences, SIFT assigns a score from 0 to 1 for nonsynonymous variants, with lower values corresponding to more harmful effects. Variants called from the whole data set (minor allele frequency $>$0.2) were annotated with SIFT. To further examine the role of purifying selection, we estimated the DFE for deleterious mutations (Eyre-Walker and Keightley 2007). The folded SFS for DFE analysis were estimated from 0- and 4-fold sites with ANGSD, and the DFE-models were fitted to these with DFE-alpha (Keightley and Eyre-Walker 2007). To account for the effects of nonequilibrium population histories, two-step N_e change was included into the models. The best fitting N_e parameters were estimated from the whole-genome SFS and fixed for module-specific analyses. A total of 1,000 nonparametric bootstrap SFS were used to define CIs for the N_es estimates.

Network Enrichment Analysis

We combined information from the coexpression network with analyses of sequence evolution to evaluate whether selection is acting differently on the coexpression modules. First, we tested whether any of the modules are overrepresented among individual genes exhibiting significant selection outlier status (F_{ST} and d_{XY} q value $<$0.01). Next, module-specific estimates of selection were acquired by summing the F_{ST} and d_{XY} estimates across genes belonging to each module and compared against one million simulated modules of equal size to assess their deviation from neutral expectations. Furthermore, we performed permutation testing by randomizing the gene-specific estimates across modules to retain selection information at individual genes while breaking up associations between them. This approach is likely to be more conservative than the simulation-based test, because the null model now includes values influenced by selection. We therefore chose a more lenient q value cutoff of 0.1 for the permutation testing. Software for conducting the network enrichment analysis is available at https://github.com/thamala/NET

Modeling of Background Selection and Genomic Features

We also assessed to what degree variation in background selection and genomic features influence our results by conducting Bayesian linear regression modeling with R package MCMCpack (Martin et al. 2011). First, to approximate the strength of background selection across the genome, we applied a model by Nordborg et al. (1996) to estimate the parameter $B = \pi/\pi_0$, where π_0 is genetic diversity in the absence of selection. Following McVicker et al. (2009) and Rettelbach et al. (2019), B at a neutral site x was estimated as:

$$B = \exp\left(-\sum_{i=1}^{n}\frac{u_d s_i}{(s_i + r_{x,i})^2}\right),$$

where u_d is deleterious mutation rate, s_i is selection against heterozygotes at site i, and $r_{x,i}$ is recombination probability between x and i. To account for selection variability at different parts of the genome, we used DFE estimated in 500-kb sliding windows to define s for the selected sites. B was then estimated for each gene by taking into account recombination rates and the number of selection targets (0-fold sites) at genes found on the same chromosome. The measure thus approximates the expected reduction in diversity at a gene due to purifying selection acting on linked sites. Next, as synonymous divergence between taxonomic groups is thought to reflect variation in mutation rates (Begun and Aquadro 1992; Cutter and Payseur 2013), we estimated the proportion of nucleotide substitutions at 4-fold sites (d_s) between the two sequenced cacao cultivars: Criollo (Argout et al. 2017), belonging to the subspecies *cacao*, and Matina 1-6 (Motamayor et al. 2013), belonging to the subspecies *sphaeorocarpum* (Motamayor et al. 2008). Besides morphological differences underlying the taxonomical grouping, these cultivars are highly differentiated genetically ($d_S = 0.02$; higher than the corresponding estimate between human and chimpanzee, e.g., Chen and Li 2001). We acknowledge, however, that these are still subspecies of cacao, so the results may not be independent of within-species variation. Orthologous gene pairs were identified with OrthoFinder (Emms and Kelly 2015), coding sequences aligned with MAFFT (Nakamura et al. 2018), and d_S estimated with an R package SeqinR (Charif and Lobry 2007). We then used weakly informative priors and 100,000 MCMC iterations, with a burn-in of 5,000

and thinning interval of 10, to fit the following model to the data:

$$y = \alpha + \beta x_B + \beta x_{\text{rec}} + \beta x_{\text{gen.den}} + \beta x_{\text{mut}} + \beta x_{\text{con}} + \beta x_{\text{exp.sum}} + \beta x_{\text{exp.var}} + \varepsilon$$

Here, y is either F_{ST}, d_{XY}, or SIFT score, α is the intercept, x_B is background selection, x_{rec} is recombination rate, $x_{\text{gen.den}}$ is gene density, x_{mut} is mutation rate, x_{con} is connectivity, $x_{\text{exp.sum}}$ is expression level, $x_{\text{exp.var}}$ is expression variance, and ε is the error term. The connectivity measure is based on pairwise correlation coefficients among RNA-Seq read counts, as estimated by WGCNA (Langfelder and Horvath 2008). All predictors were standardized to mean of 0 and SD of 1. Before running the final model, redundant predictors were removed by conducing stepwise model selection with Bayes factors. We then added a categorical predictor where the candidate modules were represented by unique identifiers and others by a common identifier. The model thus compares the modules of interest against the genome-wide background. Last, we assessed the robustness of the modeling results by randomizing the categorical predictor across genes and counting the proportion of repeats producing 95% HPD intervals more extreme than the observed ones.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

We thank D. Zhang for advice on the cacao populations, A.S. Fister for sample collection and organizational efforts, M.E. Leandro-Muñoz for sample collection, P.E. Ralph for DNA isolation, B. Epstein for advice on the coexpression network construction, E.K. Wafula for data management, and CATIE for providing access to the cacao germplasm. Computational resources were provided by the Minnesota Supercomputing Institute (MSI) at the University of Minnesota. This work was supported by the National Science Foundation (NSF) grant IOS-1546863. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

References

Argout X, Martin G, Droc G, Fouet O, Labadie K, Rivals E, Aury JM, Lanaud C. 2017. The cacao Criollo genome v2.0: an improved version of the genome for genetic and functional genomic studies. *BMC Genomics* 18(1):1–9.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al. 2000. Gene Ontology: tool for the unification of biology. *Nat Genet.* 25(1):25–29.

Barabási A-L, Oltvai ZN. 2004. Network biology: understanding the cell's functional organization. *Nat Rev Genet.* 5(2):101–113.

Barrett RDH, Hoekstra HE. 2011. Molecular spandrels: tests of adaptation at the genetic level. *Nat Rev Genet.* 12(11):767–780.

Barton NH. 1995. Linkage and the limits to natural selection. *Genetics* 140(2):821–841.

Begun DJ, Aquadro CF. 1992. Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. *Nature* 356(6369):519–520.

Beissinger T, Kruppa J, Cavero D, Ha NT, Erbe M, Simianer H. 2018. A simple test identifies selection on complex traits. *Genetics* 209(1):321–333.

Berardini TZ, Reiser L, Li D, Mezheritsky Y, Muller R, Strait E, Huala E. 2015. The Arabidopsis information resource: making and mining the "gold standard" annotated reference plant genome. *Genesis* 53(8):474–485.

Berg JJ, Coop G. 2014. A population genetic signal of polygenic adaptation. *PLoS Genet.* 10(8):e1004412.

Berg JJ, Harpak A, Sinnott-Armstrong N, Joergensen AM, Mostafavi H, Field Y, Boyle EA, Zhang X, Racimo F, Pritchard JK, et al. 2019. Reduced signal for polygenic adaptation of height in UK Biobank. *Elife* 8: e39725.

Bhatia G, Patterson N, Sankararaman S, Price AL. 2013. Estimating and interpreting F_{ST}: the impact of rare variants. *Genome Res.* 23(9):1514–1521.

Biswas S, Akey JM. 2006. Genomic insights into positive selection. *Trends Genet.* 22(8):437–446.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30(15):2114–2120.

Boyle EA, Li YI, Pritchard JK. 2017. An expanded view of complex traits: from polygenic to omnigenic. *Cell* 169(7):1177–1186.

Browning BL, Zhou Y, Browning SR. 2018. A one-penny imputed genome from next-generation reference panels. *Am J Hum Genet.* 103(3):338–348.

Buckler ES, Holland JB, Bradbury PJ, Acharya CB, Brown PJ, Browne C, Ersoz E, Flint-Garcia S, Garcia A, Glaubitz JC, et al. 2009. The genetic architecture of maize flowering time. *Science* 325(5941):714–718.

Charif D, Lobry JR. 2007. SeqinR 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. In: Structural approaches to sequence evolution. Berlin, Heidelberg: Springer. p. 207–232.

Charlesworth B. 1998. Measures of divergence between populations and the effect of forces that reduce variability. *Mol Biol Evol.* 15(5):538–543.

Charlesworth B, Morgan MT, Charlesworth D. 1993. The effect of deleterious mutations on neutral molecular variation. *Genetics* 134(4):1289–1303.

Chen F-C, Li W-H. 2001. Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. *Am J Hum Genet.* 68(2):444–456.

Chen J, Glémin S, Lascoux M. 2017. Genetic diversity and the efficacy of purifying selection across plant and animal species. *Mol Biol Evol.* 34(6):1417–1428.

Chevin LM, Hospital F. 2008. Selective sweep at a quantitative trait locus in the presence of background genetic variation. *Genetics* 180(3):1645–1660.

Chun S, Fay JC. 2011. Evidence for hitchhiking of deleterious mutations within the human genome. *PLoS Genet.* 7(8):e1002240.

Collevatti RG, Novaes E, Silva-Junior OB, Vieira LD, Lima-Ribeiro MS, Grattapaglia D. 2019. A genome-wide scan shows evidence for local adaptation in a widespread keystone Neotropical forest tree. *Heredity* 123(2):117.

Comeron JM. 2017. Background selection as null hypothesis in population genomics: insights and challenges from *Drosophila* studies. *Phiosl Trans R Soc B.* 372(1736):20160471.

Cornejo OE, Yee M-C, Dominguez V, Andrews M, Sockell A, Strandberg E, Livingstone D, Stack C, Romero A, Umaharan P, et al. 2018. Population genomic analyses of the chocolate tree, *Theobroma cacao* L., provide insights into its domestication process. *Commun Biol.* 1:167.

Cruickshank TE, Hahn MW. 2014. Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Mol Ecol.* 23(13):3133–3157.

Cutter AD, Payseur BA. 2013. Genomic signatures of selection at linked sites: unifying the disparity among species. *Nat Rev Genet.* 14(4):262–274.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, et al. 2011. The variant call format and VCFtools. *Bioinformatics* 27(15):2156–2158.

Daub JT, Hofer T, Cutivet E, Dupanloup I, Quintana-Murci L, Robinson-Rechavi M, Excoffier L. 2013. Evidence for polygenic adaptation to pathogens in the human genome. *Mol Biol Evol*. 30(7):1544–1558.

Daub JT, Moretti S, Davydov II, Excoffier L, Robinson-Rechavi M. 2017. Detection of pathways affected by positive selection in primate lineages ancestral to humans. *Mol Biol Evol*. 34(6):1391–1402.

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet*. 43(5):491–498.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29(1):15–21.

Emms DM, Kelly S. 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol*. 16(1):157.

Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa VC, Foll M. 2013. Robust demographic inference from genomic and SNP data. *PLoS Genet*. 9(10):e1003905.

Exposito-Alonso M, Vasseur F, Ding W, Wang G, Burbano HA, Weigel D. 2018. Genomic basis and evolutionary potential for extreme drought adaptation in *Arabidopsis thaliana*. *Nat Ecol Evol*. 2(2):352–358.

Eyre-Walker A, Keightley PD. 2007. The distribution of fitness effects of new mutations. *Nat Rev Genet*. 8(8):610–618.

Fisher RA. 1918. The correlation between relatives on the supposition of Mendelian inheritance. *Trans R Soc Edinb*. 53:399–433.

Gao F, Ming C, Hu W, Li H. 2016. New software for the fast estimation of population recombination rates (FastEPRR) in the genomic era. *G3 (Bethesda)* 6:1563–1571.

Garrison E, Marth G. 2012. Haplotype-based variant detection from short-read sequencing. arXiv Prepr. arXiv: 1207.3907.

Gaut BS, Seymour DK, Liu Q, Zhou Y. 2018. Demography and its effects on genomic variation in crop domestication. *Nat Plants*. 4(8):512–520.

Gossmann TI, Song BH, Windsor AJ, Mitchell-Olds T, Dixon CJ, Kapralov MV, Filatov DA, Eyre-Walker A. 2010. Genome wide analyses reveal little evidence for adaptive evolution in many plant species. *Mol Biol Evol*. 27(8):1822–1832.

Guiltinan M. 2007. Cacao. In: Biotechnology in agriculture and forestry. Berlin, Heidelberg: Springer. p. 497–518.

Hämälä T, Mattila TM, Savolainen O. 2018. Local adaptation and ecological differentiation under selection, migration and drift in *Arabidopsis lyrata*. *Evolution* 72(7):1373–1386.

Hämälä T, Savolainen O. 2019. Genomic patterns of local adaptation under gene flow in *Arabidopsis lyrata*. *Mol Biol Evol*. doi: 10.1093/molbev/msz149.

Hartfield M, Otto SP. 2011. Recombination and hitchhiking of deleterious alleles. *Evolution* 65(9):2421–2434.

He F, Arce AL, Schmitz G, Koornneef M, Novikova P, Beyer A, De Meaux J. 2016. The footprint of polygenic adaptation on stress-responsive cis-regulatory divergence in the *Arabidopsis* genus. *Mol Biol Evol*. 33(8):2088–2101.

Hermisson J, Pennings PS. 2017. Soft sweeps and beyond: understanding the patterns and probabilities of selection footprints under rapid adaptation. *Methods Ecol Evol*. 8(6):700–716.

Hoban S, Kelley JL, Lotterhos KE, Antolin MF, Bradburd G, Lowry DB, Poss ML, Reed LK, Storfer A, Whitlock MC. 2016. Finding the genomic basis of local adaptation: pitfalls, practical solutions, and future directions. *Am Nat*. 188(4):379–397.

Höllinger I, Pennings P, Hermisson J. 2019. Polygenic adaptation: from sweeps to subtle frequency shifts. *PLoS Genet*. 15(3):e1008035.

Hsieh PH, Hallmark B, Watkins J, Karafet TM, Osipova LP, Gutenkunst RN, Hammer MF. 2017. Exome sequencing provides evidence of polygenic adaptation to a fat-rich animal diet in indigenous Siberian populations. *Mol Biol Evol*. 34(11):2913–2926.

Hudson RR. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* 18(2):337–338.

Hudson RR, Kaplan NL. 1995. Deleterious background selection with recombination. *Genetics* 141(4):1605–1617.

Hudson RR, Slatkin M, Maddison WP. 1992. Estimation of levels of gene flow from DNA sequence data. *Genetics* 132(2):583–589.

Josephs EB, Berg JJ, Ross-Ibarra J, Coop G. 2019. Detecting adaptive differentiation in structured populations with genomic data and common gardens. *Genetics* 211(3):989–1004.

Josephs EB, Wright SI, Stinchcombe JR, Schoen DJ. 2017. The relationship between selection, network connectivity, and regulatory variation within a population of *Capsella grandiflora*. *Genome Biol Evol*. 9(4):1099–1109.

Keightley PD, Eyre-Walker A. 2007. Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. *Genetics* 177(4):2251–2261.

Keller SR, Levsen N, Olson MS, Tiffin P. 2012. Local adaptation in the flowering-time gene network of balsam poplar, *Populus balsamifera* L. *Mol Biol Evol*. 29(10):3143–3152.

Kim S, Lohmueller KE, Albrechtsen A, Li Y, Korneliussen T, Tian G, Grarup N, Jiang T, Andersen G, Witte D, et al. 2011. Estimation of allele frequency and association mapping using next-generation sequencing data. *BMC Bioinformatics* 12(1):231.

Kono TJY, Fu F, Mohammadi M, Hoffman PJ, Liu C, Stupar RM, Smith KP, Tiffin P, Fay JC, Morrell PL. 2016. The role of deleterious substitutions in crop genomes. *Mol Biol Evol*. 33(9):2307–2317.

Korneliussen TS, Albrechtsen A, Nielsen R. 2014. ANGSD: analysis of next generation sequencing data. *BMC Bioinformatics* 15(1):1471–2105.

Langfelder P, Horvath S. 2008. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 9(1):559.

Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv Prepr. arXiv: 1303.3997.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25(14):1754–1760.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 15(12):550.

Mähler N, Wang J, Terebieniec BK, Ingvarsson PK, Street NR, Hvidsten TR. 2017. Gene co-expression network connectivity is an important determinant of selective constraint. *PLoS Genet*. 13(4):e1006402.

Martin AD, Quinn KM, Park JH. 2011. MCMCpack: Markov chain Monte Carlo in R. *J Stat Softw*. 42(9):1–21.

Maynard Smith J, Haigh J. 1974. The hitch-hiking effect of a favourable gene. *Genet Res*. 23:23–35.

McVicker G, Gordon D, Davis C, Green P. 2009. Widespread genomic signatures of natural selection in hominid evolution. *PLoS Genet*. 5(5):e1000471.

Meisner J, Albrechtsen A. 2018. Inferring population structure and admixture proportions in low depth NGS data. *Genetics* 210(2):719–731.

Michiels A, Van den Ende W, Tucker M, Van Riet L, Van Laere A. 2003. Extraction of high-quality genomic DNA from latex-containing plants. *Anal Biochem*. 315(1):85–89.

Motamayor JC, Lachenaud P, da Silva e Mota JW, Loor R, Kuhn DN, Brown JS, Schnell RJ. 2008. Geographic and genetic population differentiation of the Amazonian chocolate tree (*Theobroma cacao* L). *PLoS One* 3(10):e3311.

Motamayor JC, Mockaitis K, Schmutz J, Haiminen N, Iii DL, Podicheti R, Zhao M, Scheffler BE, Stack JC, Feltus FA, et al. 2013. The genome sequence of the most widely cultivated cacao type and its use to identify candidate genes regulating pod color. *Genome Biol*. 14(6):1–24.

Nakamura T, Yamada KD, Tomii K, Katoh K. 2018. Parallelization of MAFFT for large-scale multiple sequence alignments. *Bioinformatics* 34(14):2490–2492.

Nei M. 1987. Molecular evolutionary genetics. New York, NY: Columbia University Press.

Nielsen R. 2005. Molecular signatures of natural selection. *Annu Rev Genet*. 39:197–218.
Nielsen R, Paul JS, Albrechtsen A, Song YS. 2011. Genotype and SNP calling from next-generation sequencing data. *Nat Rev Genet*. 12(6):443–451.
Nordborg M, Charlesworth B, Charlesworth D. 1996. The effect of recombination on background selection. *Genet Res*. 67(2):159–174.
Ossowski S, Schneeberger K, Lucas-Lledo JI, Warthmann N, Clark RM, Shaw RG, Weigel D, Lynch M. 2010. The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. *Science* 327(5961):92–94.
Pokou DN, Fister AS, Winters N, Tahi M, Klotioloma C, Sebastian A, Marden JH, Maximova SN, Guiltinan MJ. 2019. Resistant and susceptible cacao genotypes exhibit defense gene polymorphism and unique early responses to *Phytophthora megakarya* inoculation. *Plant Mol Biol*. 99(4–5):499–516.
Pritchard JK, Pickrell JK, Coop G. 2010. The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Curr Biol*. 20(4):R208–R215.
R Core Team. 2019. R: a language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing. Available from: https//www.r-project.org/.
Rettelbach A, Nater A, Ellegren H. 2019. How linked selection shapes the diversity landscape in *Ficedula* flycatchers. *Genetics* 212(1):277–285.
Reynolds J, Weir BS, Cockerham CC. 1983. Estimation of the coancestry coefficient: basis for a short-term genetic distance. *Genetics* 105(3):767–779.
Rosenberg NA, Edge MD, Pritchard JK, Feldman MW. 2018. Interpreting polygenic scores, polygenic adaptation, and human phenotypic differences. *Evol Med Public Health*. 26–34.
Salomé PA, Bomblies K, Laitinen RAE, Yant L, Mott R, Weigel D. 2011. Genetic architecture of flowering-time variation in *Arabidopsis thaliana*. *Genetics* 188(2):421–433.
Savolainen O, Lascoux M, Merilä J. 2013. Ecological genomics of local adaptation. *Nat Rev Genet*. 14(11):807–820.
Savolainen O, Pyhäjärvi T, Knürr T. 2007. Gene flow and local adaptation in trees. *Annu Rev Ecol Evol Syst*. 38(1):595–619.
Sella G, Barton NH. 2019. Thinking about the evolution of complex traits in the era of genome-wide association studies. *Annu Rev Genomics Hum Genet*. 20:1–31.
Sohail M, Maier RM, Ganna A, Bloemendal A, Martin AR, Turchin MC, Chiang CW, Hirschhorn J, Daly MJ, Patterson N, et al. 2019. Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies. *Elife* 8:e39702.
Stephan W. 2016. Signatures of positive selection: from selective sweeps at individual loci to subtle allele frequency changes in polygenic adaptation. *Mol Ecol*. 25(1):79–88.
Storey JD, Tibshirani R. 2003. Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A*. 100(16):9440–9445.
Stuart JM, Segal E, Koller D, Kim SK. 2003. A gene-coexpression network for global discovery of conserved genetic modules. *Science* 302(5643):249–255.
Supek F, Bošnjak M, Škunca N, Šmuc T. 2011. REVIGO Summarizes and visualizes long lists of gene ontology terms. *PLoS One* 6(7):e21800.
Suzek BE, Wang Y, Huang H, McGarvey PB, Wu CH. 2015. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* 31(6):926–932.
Tajima F. 1983. Evolutionary relationship of DNA sequences in finite populations. *Genetics* 105(2):437–460.
Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123(3):585–595.
Tian D, Araki H, Stahl E, Bergelson J, Kreitman M. 2002. Signature of balancing selection in *Arabidopsis*. *Proc Natl Acad Sci U S A*. 99(17):11525–11530.
Tiffin P, Ross-Ibarra J. 2014. Advances and limits of using population genetics to understand local adaptation. *Trends Ecol Evol*. 29(12):673–680.
Valluru R, Gazave EE, Fernandes SB, Ferguson JN, Lozano R, Hirannaiah P, Zuo T, Brown PJ, Leakey ADB, Gore MA, et al. 2019. Deleterious mutation burden and its association with complex traits in sorghum (*Sorghum bicolor*). *Genetics* 211(3):1075.
Vaser R, Adusumalli A, Leng SN, Sikic M, Ng PC. 2016. SIFT missense predictions for genomes. *Nat Protoc*. 11(1):1–9.
Wang J, Ding J, Tan B, Robinson KM, Michelson IH, Johansson A, Nystedt B, Scofield DG, Nilsson O, Jansson S, et al. 2018. A major locus controls local adaptation and adaptive life history variation in a perennial plant. *Genome Biol*. 19(1):72.
Watterson G. 1975. On the number of segregating sites in genetical models without recombination. *Theor Popul Biol*. 7(2):256–276.
Wright S. 1931. Evolution in Mendelian populations. *Genetics* 16(2):97–159.
Wright S. 1951. The genetical structure of populations. *Ann Eugenet*. 15:215–354.
Xie Z, Wang L, Wang L, Wang Z, Lu Z, Tian D, Yang S, Hurst LD. 2016. Mutation rate analysis via parent–progeny sequencing of the perennial peach. I. A low rate in woody perennials and a higher mutagenicity in hybrids. *Proc R Soc B*. 283(1841):20161016.
Zarrillo S, Gaikwad N, Lanaud C, Powis T, Viot C, Lesur I, Fouet O, Argout X, Guichoux E, Salin F, et al. 2018. The use and domestication of *Theobroma cacao* during the mid-Holocene in the upper Amazon. *Nat Ecol Evol*. 2(12):1879–1888.
Zhang M, Zhou L, Bawa R, Suren H, Holliday JA. 2016. Recombination rate variation, hitchhiking, and demographic history shape deleterious load in poplar. *Mol Biol Evol*. 33(11):2899–2910.
Zhou Y, Massonnet M, Sanjak JS, Cantu D, Gaut BS. 2017. Evolutionary genomics of grape (*Vitis vinifera* ssp. *vinifera*) domestication. *Proc Natl Acad Sci U S A*. 114:201709257.