Horticulture Research, 2026, **13**: uhag106
https://doi.org/10.1093/hr/uhag106
Article

Horticulture Research OXFORD

Optimizing the genomic bit budget: an information-theoretic framework for trait-enriched genotyping and stratified screening in *Theobroma cacao*

Ezekiel Ahn[1,*], Insuck Baek[2], Lalit Kandpal[2], Dapeng Zhang[1], Silvas Kirubakaran[3], Seunghyun Lim[1], Jishnu Bhatt[1], Moon S. Kim[2], Sunchung Park[1] and Lyndel W. Meinhardt[1]

[1]Sustainable Perennial Crops Laboratory, Agricultural Research Service, United States Department of Agriculture, Beltsville, MD 20705, USA
[2]Environmental Microbial and Food Safety Laboratory, Agricultural Research Service, United States Department of Agriculture, Beltsville, MD 20705, USA
[3]Grape Genetics Research Unit, Agricultural Research Service, United States Department of Agriculture, Geneva, NY 14456, USA
*Corresponding author. E-mail: ezekiel.ahn@usda.gov

Abstract

High-throughput genotyping has revolutionized horticultural breeding, yet the efficient utilization of genomic data remains a bottleneck for germplasm curation and downstream selection. Translating complex genomic information into cost-effective and readily applicable tools for clonally propagated crops requires a shift from maximizing marker density to optimizing information content. Here, we reinterpret cacao (*Theobroma cacao* L.) genotyping as an information-allocation problem and introduce an information-theoretic framework for designing minimalist, trait-enriched single nucleotide polymorphism (SNP) barcodes. Using a diverse international collection from Trinidad (ICGT) and an independent field trial in Puerto Rico (USDA-ARS Tropical Agriculture Research Station), we compress a 500+ SNP panel into a 32-marker 'CacaoCipher' barcode that preserves pairwise genetic distance structure at coarse resolution while retaining trait-aligned signal for pod index and related yield components. Barcode–space axes correlate with agronomic traits measured across environments in a limited overlap subset, supporting the portability of key signals beyond the training setting. We further quantify a heuristic 'genomic bit budget', showing how information is allocated across unique identification, ancestry structure, and trait variation. Together, this framework converts cacao germplasm from an analog collection of names into a compact digital code and provides a general template for designing low-cost, high-information marker panels for germplasm quality control and stratified screening in clonally propagated crops.

One-sentence summary

We introduce a 'genomic bit budget' framework to design a minimalist 32-SNP barcode that supports germplasm quality control, coarse ancestry screening, and trait-aligned stratification with limited cross-environment portability.

Introduction

Cacao (*Theobroma cacao* L.) stands at a critical juncture. As the economic backbone for millions of smallholders and the substrate of a multibillion-dollar global industry, the crop faces mounting pressures from climate volatility, pathogen evolution, and increasing regulatory demands for supply chain transparency. In particular, the expansion of cocoa production has been identified as a major driver of deforestation in West Africa, prompting urgent calls for rigorous traceability systems that can map products from farm to fork [1]. The sustainability of this value chain depends heavily on the efficient utilization of genetic diversity conserved in genebanks worldwide. However, mobilizing this diversity is currently constrained by an underlying challenge to genetic integrity. Historical mislabeling, synonymy (one genotype with multiple names), and homonymy (different genotypes with the same name) are pervasive in cacao germplasm collections, with error rates estimated between 15% and 30% [2, 3]. Such inaccuracies compromise breeding efficiency, contaminate seed gardens, and ultimately destabilize yield consistency in farmers' fields [4].

To resolve these ambiguities, the genomic revolution has introduced high-density single nucleotide polymorphism (SNP) arrays capable of high-precision fingerprinting, now widely adopted in the cacao community. So far, the most widely used low-density panels typically comprise <96 SNPs. Owing to their low genotyping cost and suitability for large-scale downstream applications, these panels are particularly valuable in resource-limited cocoa-producing countries. They have played a critical role in supporting on-farm research and genebank management. Comprehensive reviews of low-density SNP genotyping in cacao are provided by Wickramasuriya and Dunwell [5] and Nieves-Orduña [6]. However,

Received: 18 December 2025. **Accepted:** 12 March 2026. **Published:** 19 March 2026. **Corrected and Typeset:** 1 July 2026
Published by Oxford University Press 2026 on behalf of Nanjing Agricultural University.
This work is written by (a) US Government employee(s) and is in the public domain in the US.

Downloaded from https://academic.oup.com/hr/article/13/7/uhag106/8528763 by guest on 03 August 2026

many existing SNP panels are chosen based on heuristic criteria, without a cohesive theoretical framework to evaluate their informational efficiency or trade-offs. This limitation is evident not only in cacao but also across a range of other crops [7].

Here, we reformulate crop genotyping as a problem of information allocation. Drawing upon Shannon's information theory, we posit that genetic diversity can be quantified as entropy, and that a marker panel possesses a finite 'genomic bit budget', an information capacity that must be strategically allocated between competing biological objectives [8]. While previous studies have applied Shannon's equivocation to select markers for forensic identification [9], few have extended this concept to simultaneously optimize for population structure and agronomic utility in crop breeding. We hypothesize that a carefully curated subset of SNPs can function as a high-entropy 'digital fingerprint', capable of resolving global diversity while embedding 'agronomic intelligence' into the barcode itself.

In this context, our goal is not to replace these standard cacao genotyping panels, but rather to provide a novel complementary approach for designing more efficient barcodes that remain population-informative and can embed an agronomic signal. We tested this concept by explicitly quantifying the information capacity and allocation of a 32-SNP panel. We then evaluated minimalist barcodes alongside existing schemes using assignment performance and probability-of-identity benchmarks.

In this study, we introduce 'CacaoCipher', a framework for designing minimalist SNP barcodes from a higher density panel. Using 346 accessions from the International Cocoa Genebank, Trinidad (ICGT), we compress a 536-SNP dataset into a core set of 32 markers and evaluate the resulting barcode using distance-structure preservation, population-assignment benchmarks under class imbalance, and probability-of-identity calculations. We further construct a trait-enriched variant and evaluate trait alignment under population structure-adjusted models. For an independent environmental context, we leverage data from the USDA-ARS Tropical Agriculture Research Station (TARS) field trial in Puerto Rico and treat analyses in the ICGT–TARS overlap subset as a limited independent check of signal portability. Finally, we introduce a heuristic 'genomic bit budget' analysis to quantify how information is allocated across identity, ancestry structure, and trait variation. While developed here for *T. cacao*, this framework provides a general template for converting clonally propagated germplasm collections into compact, information-efficient codes for quality control (QC) and stratified screening.

Results

High-fidelity genomic compression via the CacaoCipher barcode

We first evaluated whether the essential genetic architecture of the diverse ICGT collection could be preserved within a minimal marker set. Compressing the initial 536-SNP dataset into the 32-SNP 'CacaoCipher' barcode resulted in a predictable reduction in absolute pairwise Hamming distances (Fig. 1). Notably, a randomly selected 32-SNP set also exhibited a comparable distance distribution, reflecting the high overall polymorphism of the source array. However, CacaoCipher was selected under explicit constraints on missingness and linkage disequilibrium (LD) redundancy, yielding a panel that is more robust for routine genebank QC than unconstrained random subsets.

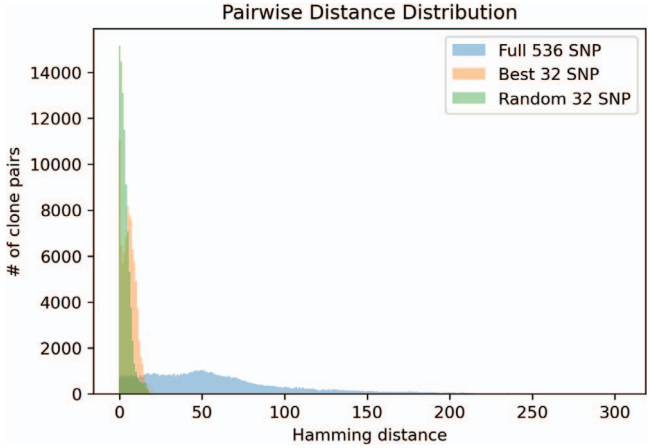

Figure 1 Distribution of pairwise Hamming distances before and after genomic compression. Histograms comparing the discriminative resolution of the full 536-SNP array versus the minimal 32-SNP CacaoCipher barcode across 346–419 ICGT accessions. While the absolute number of differences decreases with panel reduction, the 32-SNP barcode retains a broad, unimodal distribution with no collapse of interaccession distances, indicating that essential genotypic discrimination power is preserved despite a 94% reduction in marker count.

Normalizing for panel size revealed a dramatic increase in information density: the mean discriminative distance per locus rose from 0.11 in the full set to 0.18 in the 32-SNP barcode, indicating a removal of redundant signals. Crucially, this compression preserved essential population-level structure at a coarse resolution. As visualized in Fig. 2, the barcode-derived Metric Multidimensional Scaling (MDS) projection reproduced the clustering of major ancestry groups as defined by the Acc Group annotation. To quantify structure preservation beyond visualization, the 32-SNP Pure-ID panel showed a high Mantel correlation with the full genomic distance matrix ($r = 0.84$), consistent with preserved global genetic relationships. Procrustes similarity likewise supported preservation of global structure in reduced-panel embeddings (Supplementary Table S1). Code-theoretic summaries further quantified compression: the 32-SNP Pure-ID barcode resolved 321 unique code words ($\text{Log}_2 = 8.33$), whereas the trait-enriched 32-SNP variant resolved 397 unique code words ($\text{Log}_2 = 8.63$) in the observed ICGT genotype distribution (Supplementary Table S2). LD summaries indicate low-to-moderate dependence among loci (mean $r^2 \approx 0.066$ for Pure-ID 32), while total correlation confirms substantial redundancy relative to the upper-bound entropy estimate (Supplementary Table S2). This represents a substantial increase in information density relative to the full panel, achieving a high effective code rate by eliminating redundant signals.

Optimizing the cost–resolution trade-off

To empirically determine the minimal sufficient panel size, we analyzed the scaling of identification power across panel sizes $k = 8$–64. The mean Hamming distance exhibited a saturation curve, with the marginal gain in resolution diminishing sharply beyond $k = 32$ (Supplementary Fig. S1). Simulations of genotyping

Downloaded from https://academic.oup.com/hr/article/13/7/uhag106/8528763 by guest on 03 August 2026

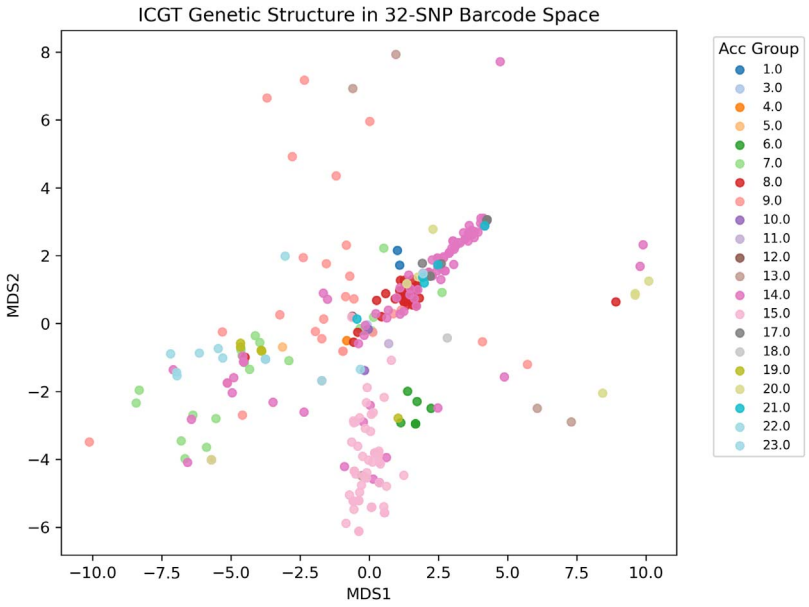

Figure 2 Recapitulation of population structure in the minimal barcode space. MDS projection of the ICGT collection based on pairwise Hamming distances derived from the 32-SNP barcode. Accessions are colored by their Acc Group label (modern 11-group genetic classification). The distinct clustering of ancestral groups confirms that the minimal barcode faithfully encodes the broad phylogeographic structure of the germplasm, mirroring results from high-density arrays.

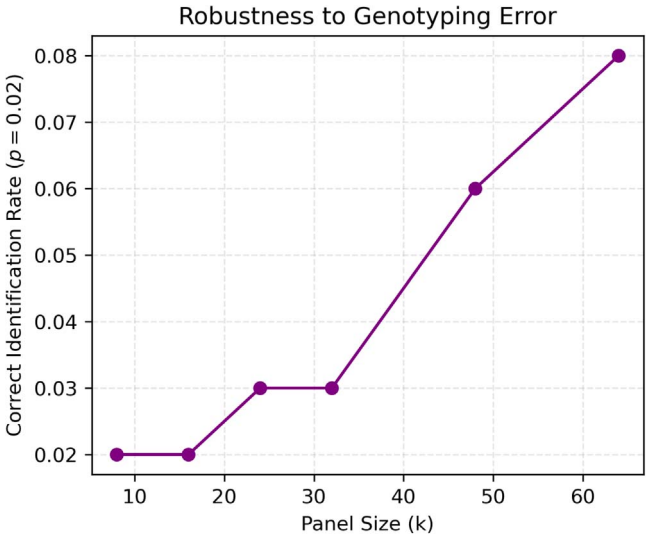

Figure 3 Robustness of identification against genotyping errors. Simulation of misidentification probability across varying panel sizes (k) and per locus genotyping error rates (0%–5%). While smaller panels ($k < 24$) show rapid degradation in accuracy, the 32-SNP panel maintains a correct identification rate exceeding 95% even at a 5% error rate, demonstrating suitability for low-cost, routine genotyping platforms.

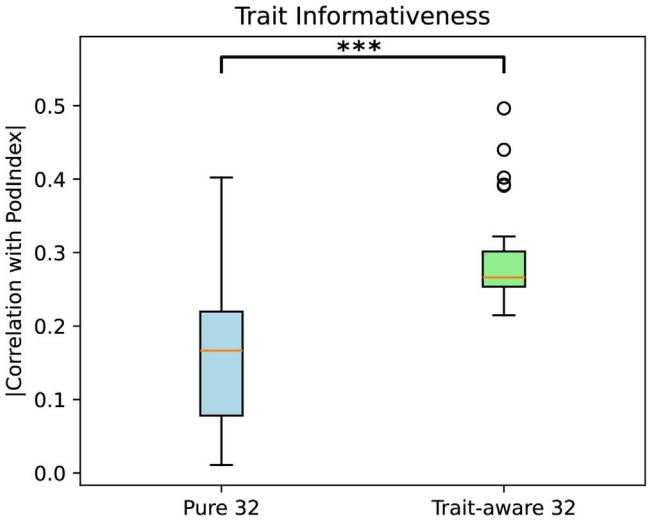

Figure 4 Trade-off between genotypic discrimination and agronomic information. Comparison of panel performance metrics for the pure identification (ID) panel versus the trait-enriched panel. Boxplots display the distribution of marker-trait correlations ($|r|$) for Pod Index. The trait-enriched design achieves a significant increase in phenotypic association (median $|r|$ shift) with only a marginal reduction in mean Hamming distance, validating the dual-purpose utility of the barcode. Asterisks indicate statistical significance from a one-sided two-sample t-test ($^{***}P < 0.001$).

error demonstrated that the 32-SNP threshold serves as a stability tipping point: while smaller panels (k <24) were highly sensitive to error, the 32-SNP barcode maintained a correct identification probability > 95% even at per locus error rates of 5% (Fig. 3). Together, these analyses suggest that 32 SNPs provide a practical knee point that balances genotyping cost against robustness for routine identity verification, while acknowledging that higher density panels remain preferable for applications requiring pedigree-level resolution.

Enhancing agronomic utility through trait-enriched design

Moving beyond random marker selection, we engineered a trait-enriched variant of the 32-SNP panel designed to capture yield-related signals. Compared to a purely identity-focused panel, the trait-enriched barcode accepted a marginal 7% reduction in

Downloaded from https://academic.oup.com/hr/article/13/7/uhag106/8528763 by guest on 03 August 2026

Table 1 Comparative performance of Pure-ID versus Trait-enriched barcodes.

Biological category	$\|r\|$ ID-only 32	$\|r\|$ Trait-enriched 32	% gain in $\|r\|$	Distance reduction (%)
Pod efficiency/yield	0.158	0.22	39.1	7.1
Wet bean mass (yield)	0.173	0.213	23.3	5.1
Fruit surface anthocyanin intensity (mature ridges)	0.149	0.246	65.2	13.1
Flower morphology (style length)	0.144	0.213	48.4	8.7

Mean marker–trait correlation coefficients ($|r|$) for key agronomic traits (Pod index, wet bean mass, fruit surface anthocyanin intensity, and flower morphology) comparing the random/pure-ID panel selection against the trait-enriched algorithm. Percentage gains indicate the enhanced predictive value embedded in the trait-enriched barcode.

Table 2 Population assignment performance under the 11-group cacao classification.

Panel	Markers (k)	Overall accuracy	Balanced accuracy	Macro-F1
Pure-ID 32-SNP	32	0.628 (0.602–0.655)	0.451 (0.418–0.484)	0.419 (0.388–0.449)
Random 32-SNP	32	0.550 (0.515–0.585)	0.363 (0.330–0.396)	0.333 (0.303–0.363)
Pure-ID 96	96	0.674 (0.654–0.693)	0.428 (0.398–0.457)	0.415 (0.384–0.446)

Mean performance with 95% confidence intervals across repeated stratified cross-validation (five folds × five repeats). Metrics include overall accuracy, balanced accuracy, and macro-F1. The random 32-SNP baseline was generated by sampling markers uniformly at random within each fold.

mean genetic distance (5.65–5.25) in exchange for substantial gains in phenotypic association. Specifically, the mean marker–trait correlation ($|r|$) increased by 39% for Pod Index and 65% for fruit surface anthocyanin intensity (Fig. 4, Table 1). Beyond yield, the trait-enriched design also enhanced associations for morphological markers such as style length (48% gain), demonstrating the panel's versatility across diverse biological traits. This strategic trade-off demonstrates that agronomic intelligence can be embedded into a fingerprinting panel without compromising its primary identification function. To assess whether these signals extend beyond ancestry-driven confounding, we evaluated barcode-derived coordinates under population structure-adjusted models. Under principal component (PC)-adjusted models, the trait-enriched 32-SNP panel retained a residual association with Pod Index (robust $P = 0.018$; stratified permutation $P < 0.001$), whereas the identity-focused panel's association was attenuated after adjustment. PC-adjusted and stratified-permutation results are summarized in Supplementary Table S3. Consistent with ancestry-confounding expectations, the Pure-ID panel's Pod Index association was not significant after PC adjustment ($P = 0.477$), whereas the trait-enriched panel retained a residual signal (Supplementary Table S3).

Population informativeness under the 11-group cacao framework

Given strong class imbalance (including rare groups with very few samples), we report balanced accuracy alongside overall accuracy. Using repeated stratified cross-validation (five folds repeated five times), the 32-SNP Pure-ID panel achieved a mean balanced accuracy of 0.451, outperforming random 32-SNP baselines (balanced accuracy 0.363) and approaching the performance of a higher density 96-marker Pure-ID baseline panel (Table 2). While the panel demonstrated a nontrivial assignment signal relative to random selection, it did not fully resolve all 11 groups, particularly for underrepresented classes. Accordingly, we interpret the 32-SNP panel as enabling coarse-to-intermediate ancestry screening for germplasm curation rather than definitive multiclass assignment. Across the major-group subset ($n \geq 5$), misclassifications were not uniformly distributed and were often concentrated into the most prevalent Acc Group, consistent with class-imbalance effects (Supplementary Fig. S2). Fold-wise performance distributions are shown in Fig. 5, highlighting that the Pure-ID 32-SNP panel consistently exceeds the random baseline across repeated splits.

Discriminative power and probability of identity

To benchmark the forensic resolution of the minimalist barcode, we calculated multilocus probabilities of identity following Waits *et al*. [10] (Table 3). Under Hardy–Weinberg assumptions, the full 536-SNP panel (522 informative loci) yielded an extremely low probability of identity among unrelated individuals ($\log_{10} P(ID)_{\text{unrelated}} \approx -140.6$) and full siblings ($\log_{10} P(ID)_{sibs} \approx -57.9$), effectively eliminating the risk of confusing distinct genotypes. The 32-SNP Pure-ID barcode, while less extreme, still achieved $\log_{10} P_{ID,\text{unrelated}} \approx -8.4$ ($\approx 4 \times 10^{-9}$), corresponding to an expected collision rate of roughly 1 in 2.6×10^8 unrelated individuals. In contrast, $\log_{10} P_{ID,sibs} \approx -3.6$ indicates that the panel is not sufficiently conservative for robust parentage or sibship reconstruction. Thus, our 32-SNP design is well suited for identity verification and population assignment in germplasm curation, but, consistent with current community practice, denser panels such as the 96-SNP working-group standard remain preferable for likelihood-based pedigree analysis.

Cross-environmental stability of barcode-predicted traits

To validate the barcode's predictive utility outside the training environment, we projected 24 accessions shared between the

Downloaded from https://academic.oup.com/hr/article/13/7/uhag106/8528763 by guest on 03 August 2026

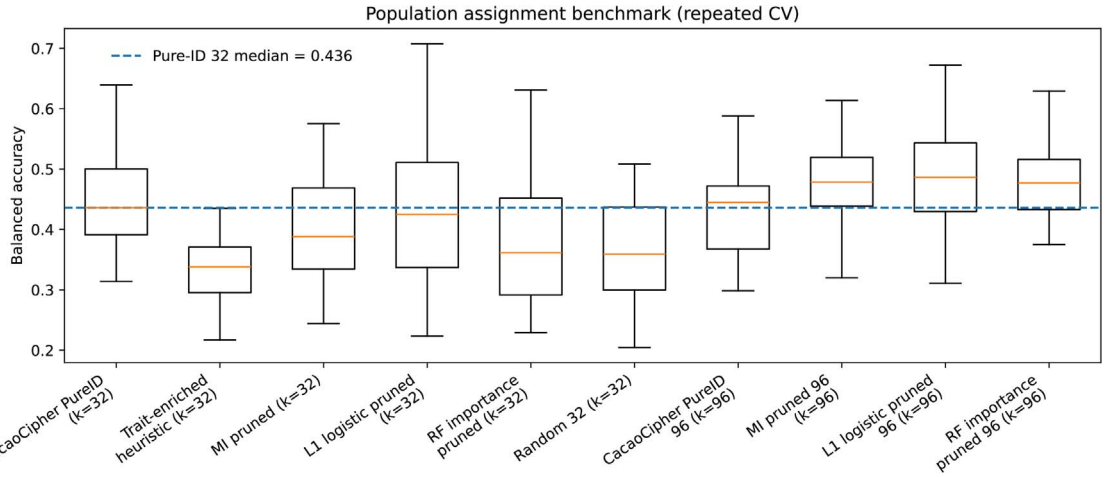

Figure 5 Population assignment benchmark under repeated stratified cross-validation. Fold-wise distributions of balanced accuracy across repeated stratified cross-validation (five folds × five repeats) comparing the 32-SNP Pure-ID panel against a random 32-SNP baseline (and additional benchmark panels where indicated). The minimalist panel exceeds random selection but shows reduced resolution for underrepresented classes, consistent with coarse-to-intermediate ancestry screening utility. A dashed horizontal line indicates the median balanced accuracy of the 32-SNP Pure-ID panel across folds and repeats (note that Table 2 reports mean performance).

Table 3 Multilocus probability of identity ($P(ID)$) for unrelated individuals and full siblings.

Panel	Markers used for $P(ID)$ (k)	$\log_{10}P(ID)_{\text{unrelated}}$	$\log_{10}P(ID)_{sibs}$
Full 536-SNP set	522	−140.59	−57.94
Pure-ID 32-SNP barcode	32	−8.42	−3.62

Values are reported on a $\log_{10}$ scale following Waits *et al*. [10]. Lower values indicate higher discriminative power.

Table 4 Cross-environmental correlation matrix.

Trait	Biological category	r (MDS1)	r (MDS2)
Yield	Productivity	0.488	−0.086
Total number of pods	Productivity	0.567	0.279
Pod index	Pod efficiency	0.047	0.723
Dry weight	Seed size/quality	−0.116	−0.635
Fresh weight	Seed size/quality	−0.346	−0.284
Number of seeds	Reproductive output	0.118	−0.162
Pod weight	Pod/fruit biomass	−0.316	−0.321

Pearson correlation coefficients (r) between the two principal barcode axes (MDS1, MDS2) derived from ICGT genotypes and various phenotypic traits measured at the TARS field site. Traits with significant correlations ($P < 0.05$) highlight the environmental stability of the genetic signals encoded by the barcode.

Trinidad (ICGT) and Puerto Rico (TARS) collections into the barcode space. Despite environmental differences, the genetic coordinates derived from Trinidad showed significant correlations with field performance in Puerto Rico. The first barcode dimension (MDS1) correlated positively with Yield ($r \approx 0.49$) and Total Pods ($r \approx 0.57$), while the second dimension (MDS2) captured a trade-off between Pod Index ($r \approx 0.72$) and Dry Bean Weight ($r \approx -0.64$) (Fig. S3, Table 4). These correlations suggest that barcode-derived coordinates capture trait-aligned gradients that can exhibit cross-environment consistency in the limited ICGT–TARS overlap subset ($n = 24$), rather than reflecting purely identity-based variation.

Barcode-based quality control and rationalization

Beyond identification and prediction, we applied the 32-SNP panel to assess redundancy within the ICGT collection. By scanning for pairs with near-identical barcodes (Hamming distance ≤ 1), we identified several large clusters of synonymous accessions (Supplementary Data S1). The largest cluster, rooted at core 'NA326', contained 148 accessions, suggesting significant historical duplication or mislabeling within the collection. Notably, this single synonym group accounts for over one-third (35%) of the entire genotyped collection, underscoring the

Downloaded from https://academic.oup.com/hr/article/13/7/uhag106/8528763 by guest on 03 August 2026

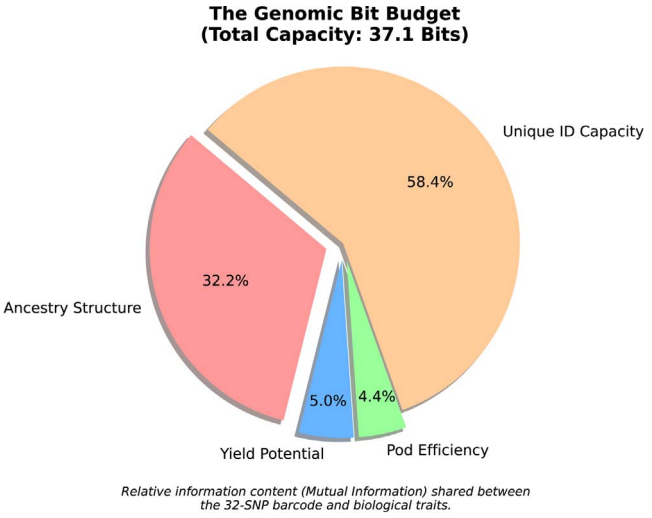

Figure 6 The genomic bit budget: strategic information allocation. Pie chart illustrating a heuristic decomposition of the 32-SNP barcode's total information capacity (∼37 bits; estimated by summing per locus Shannon entropies of genotype states under approximate independence) into components associated with Ancestry Structure (∼32%), Yield Potential & Pod Efficiency (∼10%), and Unique Identification (∼58%), using mutual information to quantify these associations. Because targets are correlated, the MI-based budget should be interpreted as a heuristic allocation rather than a strict orthogonal decomposition. Note that per locus entropy may exceed 1 bit because genotype states are multilevel (0/1/2), so total capacity can exceed the number of loci.

massive redundancy currently managed in field genebanks. This result highlights the panel's immediate utility for rationalizing genebank holdings and reducing maintenance costs.

The genomic bit budget: quantifying information allocation

Finally, we quantified the information content of the CacaoCipher barcode using complementary capacity summaries, including realized discriminative capacity from observed unique code words (log2(n)) and an upper-bound capacity (∼37 bits) obtained by summing per locus entropies under approximate independence. LD redundancy and realized capacity summaries are reported in Supplementary Table S2. Because genotypes are multilevel (e.g. 0/1/2), per locus Shannon entropy can exceed 1 bit; thus, summed entropies can exceed the number of loci and are interpreted here as an upper-bound capacity. Visualizing the 'digital fingerprints' for high- vs low-yielding clones (Supplementary Fig. S4) illustrates the contrasting patterns encoded by the barcode. Decomposing this capacity via mutual information analysis revealed a specific 'bit budget' allocation: ∼32% (11.9 bits) encodes Ancestry Structure, ∼10% (3.5 bits) captures Yield and Pod Efficiency potential, while the remaining ∼58% (21.7 bits) serves as the reservoir for Unique Identification (Fig. 6). This breakdown supports the view that the barcode acts as a multipurpose digital key, simultaneously serving ancestry, yield, and identity-related objectives.

Discussion

Minimal barcodes as efficient genomic codes

In this study, we revisited an extensively characterized cacao germplasm panel [11] to address a key challenge in resource-limited breeding: optimizing the trade-off between genotyping cost and information gain. We demonstrated that a carefully chosen 32-SNP barcode, 'CacaoCipher', preserves major pairwise genetic structure at a coarse resolution suitable for routine germplasm QC and ancestry screening while dramatically increasing information efficiency per locus. This aligns with recent efforts in other crops, such as the development of a 20-SNP panel for *Cannabis sativa* cultivar differentiation [12] and QC panels in maize [13, 14], consistent with a broader shift toward minimal, high-throughput genotyping standards.

Crucially, our approach demonstrates that substantial redundancy can be removed from current genotyping panels without collapsing the underlying genetic structure. This suggests that once the full genotype matrix is mapped via high-density genotyping or sequencing [15], genebanks can transition to highly compressed, standardized panels for routine identity checks, significantly reducing the 'molecular overhead' of germplasm management.

Embedded agronomic signal within a fixed marker budget

Hereafter, we refer to this design as a trait-enriched barcode (i.e. a panel selected to increase trait alignment while controlling for population structure). A second key finding is that a trait-enriched 32-marker budget can carry nontrivial agronomic information without sacrificing its utility for identification. By modifying the panel selection objective to reward SNP–trait association (functional markers), we achieved consistent gains in correlations with pod index and bean weight. Functional markers are increasingly used to connect low-density genotyping to agronomic targets, although dense genome-wide marker sets remain the standard for high-accuracy selection [16]. However, such a trait-targeted design is still rarely integrated into minimal QC-oriented panels. Importantly, the Pod Index signal for the trait-enriched panel persists after population-structure adjustment, whereas the Pure-ID panel's association attenuates (Supplementary Table S3). To improve interpretability, we provide available array-based gene annotations for markers in the final panels in the Supplementary Data S1 workbook, noting that these annotations are descriptive and do not imply causal trait loci.

Our results demonstrate that 'identity' and 'prediction' are not mutually exclusive design goals but define a continuum. While functional markers have been used in low-density genotyping of cacao germplasm [17], our framework demonstrated that 'identity' and 'prediction' could be operationalized into a unified tool. A genebank may favor a pure ID-optimized barcode to maximize discrimination [18], whereas a breeding program may prefer a trait-enriched barcode that reallocates a portion of its bit budget toward yield-related variation. This dual-purpose utility offers a practical advantage over static marker sets, allowing curators to navigate the trade-off explicitly.

Downloaded from https://academic.oup.com/hr/article/13/7/uhag106/8528763 by guest on 03 August 2026

Model portability and environmental stability: validating the barcode in Puerto Rico

A critical limitation of many genomic markers is their inconsistency across diverse environments. However, our results suggest that the barcode coordinates derived from the Trinidad germplasm successfully correlated with agronomic traits (e.g. pod index, yield) measured in an independent Puerto Rican field trial. This cross-environment comparison provides a limited but independent check of signal portability. It suggests that the trait-enriched barcode captures biological signals that persist despite substantial environmental variation. Because the ICGT–TARS overlap is modest ($n = 24$), these results should be interpreted as supportive evidence rather than a definitive external validation of prediction performance.

Specifically, the barcode axes derived in Trinidad showed consistent associations with yield traits in Puerto Rico, indicating that the selected markers are tracking stable agronomic gradients rather than local environmental noise. This aligns with findings by Ramirez-Ramirez *et al*., who demonstrated high congruence between marker systems across populations, reinforcing the transferability of well-designed genomic signals [19]. By validating the barcode in a distinct environment without retraining, we provide evidence that the 'Genomic Bit Budget' optimized in one genebank can serve as a predictive baseline for another, facilitating global data integration.

A practical tool for germplasm curation and identity management

From an operational perspective, the 32-SNP barcode emerges as a critical tool for quality control. Mislabeling and synonymy are pervasive issues in clonal crops, with error rates often exceeding 30% in historical collections [4, 20]. Our distance-based analysis flagged significant redundancies and potential labeling errors, mirroring findings in Colombian cacao banks where genetic redundancy reached ∼53% [21]. Similar taxonomic erosion has been reported in Brassica genebanks [22], underscoring the universal need for molecular auditing. From a numerical standpoint, the multilocus probability of identity for unrelated individuals under our 32-SNP panel ($\log_{10} P_{ID,\text{unrelated}} \approx -8.4$) is comparable to, or better than, many SSR-based fingerprinting schemes that are routinely deployed for clonal crops [14, 20]. This level of resolution is more than adequate for detecting label collisions among unrelated accessions and for flagging large synonym clusters, as evidenced by the NA326-centered group that accounts for over one-third of the ICGT collection. However, the substantially higher $P(ID)_{sibs}$ values relative to dense SNP panels reinforce that minimal barcodes should complement, rather than replace, high-density arrays in applications that require unambiguous parentage inference. In practice, some platforms price assays per 96-well plate, which can narrow the marginal cost difference between 32 and 96 markers. Even so, 32-SNP panels remain advantageous for flexible targeted genotyping (e.g., KASP/amplicon/nanofluidic assays), rapid QC deployment, and reduced downstream data handling.

By deploying 'CacaoCipher', curators can objectively triage accessions: retaining representatives from large synonym clusters and focusing resources on genuinely distinct material. This data-driven rationalization is essential for identifying 'off-types' or pollen contamination that can disastrously impact variety recommendations [4]. Furthermore, the ability to authenticate distinct clones with a minimal panel has direct forensic applications in the high-value cocoa supply chain, enabling verification of single beans [23] or premium varietal products.

The genomic bit budget: from locus lists to codes

Perhaps the most conceptual contribution of this work is the treatment of a SNP panel as a digital code with a finite 'bit budget'. By computing the Shannon entropy of each marker, we quantified the total information capacity of the barcode. This information-theoretic approach has theoretical precedents in forensic genetics, where Shannon's equivocation is used to select Y-STR markers that maximize discriminatory power [9], and in genetic mapping, where entropy-based SNP selection has been shown to reduce genotyping load by up to 30% [24].

We extend this framework to decompose the barcode's capacity into components devoted to ancestry, identity, and traits. In our case, the barcode exhibits a total entropy of ∼37 bits, with distinct fractions allocated to population structure and agronomic performance. This echoes the use of mutual information to detect gene–phenotype associations in complex diseases [25, 26]. Consistent with this framing, LD redundancy and total correlation indicate substantial nonindependence among loci, reinforcing that summed per locus entropy is an upper-bound capacity measure (Supplementary Table S2). Framed this way, panel design becomes an exercise in information allocation. Instead of asking 'Which markers are statistically significant?', we ask 'How should we spend our limited bits?' This framing provides a rigorous way to discuss trade-offs between forensic resolution and agronomic prediction.

Future directions

While our 32-SNP barcode efficiently captures population structure and key trait associations, it is inherently conditioned on the original SNP discovery platform. Future efforts could employ reference-optional pipelines to incorporate novel variants from wild populations. Nevertheless, the 'Genomic Bit Budget' framework presented here offers a generalizable path for clonal crops: moving from static marker lists to information-efficient codes that optimize the trade-off between forensic resolution and agronomic prediction. This shift toward trait-enriched identity establishes a new standard for cost-effective, readily applicable germplasm management in cacao and beyond. We did not experimentally evaluate performance under traceability conditions such as degraded DNA, mixed-origin samples, or platform-specific amplification bias, and thus do not claim validated adulteration detection in this study. A practical next step is to benchmark the 32-SNP panels using controlled DNA degradation series and synthetic mixture titrations across common targeted genotyping platforms (e.g. KASP/amplicon/nanofluidic assays). These tests would quantify robustness and detection limits while keeping the current framework focused on germplasm QC and stratified screening.

Downloaded from https://academic.oup.com/hr/article/13/7/uhag106/8528763 by guest on 03 August 2026

Materials and methods

Plant materials and phenotypic resources

To construct and validate the minimalist barcode framework, we integrated genomic and phenotypic data from two distinct international germplasm collections. The primary training population consisted of 346 *T. cacao* accessions from the ICGT, representing a broad spectrum of genetic groups as annotated by the Acc Group labels in Bekele *et al*. [11]. Phenotypic data for this panel included extensive morphological descriptors and yield components. Crucially, yield potential was assessed using Pod Index, defined as the number of pods required to produce 1 kg of dried beans (1000/ (avg.dried bean mass × avg.seed number)), serving as an inverse metric where lower values indicate higher efficiency [11]. Genotypic data comprised 536 high-quality SNPs derived from a 6 K SNP array, filtered for call rates >95% and minor allele frequency (MAF) >0.05 [27]. For population-level benchmarking, the ICGT accessions were assigned to an 'Acc Group' label reflecting genetic group metadata provided with the collection. Sample sizes vary across analyses due to missingness in Acc Group labels and trait measurements; accordingly, some analyses use the full set of up to 419 genotyped ICGT accessions, whereas others use the fully phenotyped training subset ($n = 346$) or the ICGT–TARS overlap subset ($n = 24$).

For cross-environmental validation, we utilized an independent dataset from the TARS in Puerto Rico, comprising 173 accessions evaluated under humid tropical conditions [28]. To provide environmental context for yield stability analysis, historical daily temperature and precipitation records for the evaluation period were retrieved from the NOAA Global Historical Climatology Network-Daily (GHCN-Daily) database [29]. Weather records were used solely to contextualize trial conditions and were not used as predictive covariates in the barcode models. A subset of 24 accessions common to both ICGT and TARS was identified via taxonomic standardization and used to assess the stability of barcode-predicted agronomic traits across environments.

Information-theoretic framework and 'bit budget' analysis

All data processing, information-theoretic calculations, and statistical analyses were performed in Python (v3.11) using the scikit-learn [30] and pandas libraries. We reformulated the marker selection process as an information allocation problem, quantifying genetic diversity as entropy (Sherwin, 2010). The information capacity of the SNP panel was calculated using Shannon entropy (H). For a given marker m with genotype frequencies $p(x)$, the entropy was defined as:

$$H(m) = -\sum P(x)\log_2 p(x)$$

The total information capacity of the panel (H_{total}) was approximated by the sum of individual marker entropies, providing a theoretical upper bound on discriminative power under an assumption of approximate marker independence [31]. Because ultraminimal panels inevitably retain LD and other dependencies, we additionally quantified redundancy within each candidate panel using pairwise LD (r^2) summaries and an information-theoretic dependence metric (total correlation). We further reported realized discriminative capacity by the number of observed unique barcode code words and its $\log_2$ transform, which provides an empirical lower bound proxy for effective capacity under the observed sampling distribution.

For clarity, we distinguish an empirical realized capacity summarized by the number of observed unique barcode code words (and $\log_2$ of this value) from the theoretical upper bound capacity obtained by summing per locus entropies (∼37 bits). To obtain a heuristic 'genomic bit budget', we computed the mutual information between the 32-SNP barcode and three categorical targets (ancestry group, yield class, pod-efficiency class) and summed these contributions. Because these targets are themselves correlated, this bit budget should be viewed as an approximate mutual information (MI)-based partition of H_{total} rather than a strict orthogonal decomposition of entropy. Accordingly, MI terms were interpreted as pragmatic bookkeeping signals for panel design rather than evidence of causal trait loci. The remaining portion of H_{total} ($H_{\text{total}} - \sum I$) was interpreted as capacity primarily devoted to unique identification.

Construction of minimal trait-enriched barcodes

We treated the marker selection not as a random filtering process, but as a mathematical optimization problem. By calculating Shannon Entropy (H) and Mutual Information (I) for each locus, we identified a minimal set that maximizes discriminatory power while minimizing redundancy. This heuristic greedy procedure provides a reproducible, practical solution that balances discrimination and trait informativeness under explicit constraints (missingness and LD pruning). To achieve this, we implemented a stepwise, greedy forward-selection procedure to compress the initial 536-SNP set into a minimal 32-SNP panel, designated as 'CacaoCipher'. For benchmarking analyses that compare alternative panel construction strategies (e.g. mutual information, L1-regularized logistic regression, and random panels), marker selection was performed strictly within each training fold of cross-validation, and the resulting marker set was then evaluated on the corresponding held-out fold to prevent information leakage. The algorithm optimized a composite objective function designed to maximize genotypic discrimination while retaining agronomic information [7]. The scoring function for marker selection was defined as:

$$Score = \alpha \cdot \text{Distance}(S) + \beta \cdot \text{Association}(S)$$

where Distance(S) represents the mean pairwise Hamming distance of the selected subset S, and Association (S) represents a heuristic association term computed as the mean absolute Pearson correlation ($|r|$) between candidate markers and ICGT training traits (pod index, wet bean mass), used only as a panel-design signal rather than as evidence of causal loci. For the identification-only panel, we set $\beta = 0$; for the trait-enriched panel, we set $\alpha = \beta = 1$. The robustness of the selected panels was evaluated by simulating genotyping errors via Monte Carlo simulation (1000 replicates), quantifying the decay of identification accuracy at error rates ranging from 0% to 10% [32].

Downloaded from https://academic.oup.com/hr/article/13/7/uhag106/8528763 by guest on 03 August 2026

Multilocus probability of identity

To benchmark the discriminative power of the minimalist panel against established cacao genotyping standards, we calculated the multilocus probability of identity ($P(ID)$) for unrelated individuals and full siblings, following the approach of Waits *et al*. [10]. For each SNP, allele frequencies were estimated from the ICGT panel, and locus-wise $P(ID)$ values were computed under Hardy–Weinberg equilibrium. Multilocus $P(ID)$ for unrelated individuals was obtained as the product of locus-wise $P(ID)$ values across all markers in a panel. The sibling probability of identity $P(ID)_{sibs}$ was derived using the standard approximation that incorporates both homozygote and heterozygote matching probabilities at each locus [10].

For the 'full' panel, only loci with nonmissing genotypes across the vast majority of accessions were retained for $P(ID)$ calculations, resulting in 522 informative markers out of the initial 536. For the 32-SNP Pure-ID barcode, all selected loci were included. Final values are reported in $\log_{10}$ scale to facilitate comparison across panels that differ in marker number. $P(ID)$ was computed under Hardy–Weinberg assumptions and should be interpreted as an approximate benchmark in a structured germplasm panel.

Cross-environmental validation of barcode-predicted traits

To validate the barcode's utility beyond the training environment, ICGT-derived barcode coordinates were projected into a low-dimensional space using MDS [33]. These genetic coordinates were then compared against independent field performance data from the Puerto Rico (TARS) trial to assess whether panel-derived trait-aligned signals exhibited cross-environment consistency [4]. Because only a subset of accessions overlapped between collections, this validation was treated as a limited but independent check rather than a definitive external test.

Population assignment benchmarking and structure preservation under the modern 11-group framework

To evaluate whether the minimalist barcode remained informative with respect to the current cacao population framework, we assessed population assignment performance under the modern 11-group classification implemented in the international germplasm community [11, 34]. We used the 'Acc Group' label from the Bekele *et al*. dataset as the target variable and restricted analyses to accessions with nonmissing group assignments [11]. Genotype matrices were constructed either from the full 536-SNP set or from the 32-SNP Pure-ID panel. Missing genotypes were handled using a deterministic encoding scheme applied consistently across cross-validation splits to preserve panel dimensionality. Specifically, the same marker set and encoding rules were applied to both training and held-out folds, ensuring identical feature dimensionality across splits. For benchmarking, we also constructed a 96-marker Pure-ID baseline by extending the same unsupervised Pure-ID selection procedure used for the 32-marker panel to $k = 96$ (heterozygosity ranking with missingness filtering and correlation pruning). This panel serves as a matched, higher density comparator within the same SNP array dataset and is not intended to represent a community-wide standardized 96-SNP panel.

Population assignment was carried out using a Random Forest classifier implemented in scikit-learn (n_estimators = 500; default Gini impurity) [30]. We used repeated stratified cross-validation (five folds repeated five times). For each held-out split, we recorded overall accuracy, balanced accuracy (mean of per class recalls), and macro-F1 to account for the pronounced class imbalance, as several genetic groups were represented by very few accessions. To ensure valid stratification when rare classes were present, labels with fewer than the number of folds were collapsed into an 'Other' category prior to splitting. Performance summaries are reported as fold-wise distributions and as mean values with 95% confidence intervals. Confidence intervals were computed empirically from the distribution of fold-wise performance values across repeats. Given the severe class imbalance, we additionally report fold-aggregated, row-normalized confusion matrices for the subset of major Acc Groups ($n \geq 5$) to visualize class-specific error patterns.

In addition to supervised classification metrics, we assessed unsupervised structure preservation by comparing the pairwise genetic distance matrix derived from the full SNP set against the corresponding distance matrix derived from each reduced panel. Similarity was quantified using Mantel correlation (Pearson and Spearman) and Procrustes similarity between low-dimensional embeddings.

Trait–ancestry confounding control and within-group analysis

To disentangle trait-associated signal from ancestry-driven confounding, we evaluated panel-derived coordinates against traits under both unadjusted and population structure-adjusted models. Structure adjustment was performed by including the top principal components derived from the full SNP set as covariates. Regression inference used heteroskedasticity-robust standard errors. To further guard against inflated significance under structured labels, we computed empirical P-values using a stratified permutation that shuffles trait values within genetic groups. Finally, we computed within-group correlations for major groups with sufficient sample size to assess whether associations persisted beyond between-group differences.

Acknowledgements

Mention of any trade names or commercial products in this article is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the US Department of Agriculture. USDA is an equal opportunity provider and employer, and all agency services are available without discrimination. This work was supported by the US Department of Agriculture, Agricultural Research Service, In-House Projects Nos. 8042-21220-258-000-D and 8072-41000-113-000-D.

Author contributions

E.A. conceived and designed the study, supervised the project, curated the data, performed the formal analyses, and wrote the

Downloaded from https://academic.oup.com/hr/article/13/7/uhag106/8528763 by guest on 03 August 2026

original draft of the manuscript. I.B. contributed to software development, validation, and formal analyses, and revised the manuscript. L.K. and J.B. provided resources and contributed to manuscript editing. S.K., S.L., and S.P. contributed to methodology development, data analysis, and manuscript editing. M.S.K. and L.W.M. provided supervision, resources, and funding acquisition and contributed to manuscript revision. All authors read and approved the final manuscript.

Data availability

The complete Python code used to run the analysis pipeline is publicly available at the project GitHub repository: https://github.com/EJSAHN/cacaocipher-analysis-pipeline. Raw/compiled input spreadsheets used to build the local data_raw/directory are not redistributed in the repository because they include third-party materials and/or curated working files assembled from multiple sources. Instead, the repository documents the original data sources and provides step-by-step instructions for obtaining the inputs directly from the primary sources and reproducing the analyses locally.

Portions of the codebase were refactored with assistance from an AI-based coding tool (ChatGPT) to improve readability and enable public release; the underlying analysis logic was not intentionally altered. The deposited manuscript and supplementary materials should be treated as the authoritative record for final reported values.

Conflicts of interest statement

The authors declare no conflicts of interest.

Supplementary material

Supplementary material is available at *Horticulture Research* online.

References

1. Renier C, Vandromme M, Meyfroidt P. *et al.* Transparency, traceability and deforestation in the Ivorian cocoa supply chain. *Environ Res Lett*. 2023;18:024030
2. CacaoNet. A global strategy for the conservation and use of cacao genetic resources, as the foundation for a sustainable cocoa economy. Montpellier, France: Bioversity International, 2012
3. Turnbull C, Butler DR, Cryer NC. *et al.* Tackling mislabelling in cocoa germplasm collections. *INGENIC Newsl*. 2003;8:8–11
4. Padi FK, Ofori A, Takrama J. *et al.* The impact of SNP fingerprinting and parentage analysis on the effectiveness of variety recommendations in cacao. *Tree Genet Genomes*. 2015;11:44
5. Wickramasuriya AM, Dunwell JM. Cacao biotechnology: current status and future prospects. *Plant Biotechnol J*. 2018;16:4–17
6. Nieves Orduña HE. *Genetic Diversity of Cacao (Theobroma Cacao L.) Populations and Agronomic Traits for Breeding*. PhD thesis, University of Göttingen, Göttingen, Germany. 2023
7. Dou T, Wang C, Ma Y. *et al.* CoreSNP: an efficient pipeline for core marker profile selection from genome-wide SNP datasets in crops. *BMC Plant Biol*. 2023;23:580
8. Sherwin WB. Entropy and information approaches to genetic diversity and its expression: genomic geography. *Entropy*. 2010;12:1765–98
9. Siegert S, Roewer L, Nothnagel M. Shannon's equivocation for forensic Y-STR marker selection. *Forensic Sci Int Genet*. 2015;16:216–25
10. Waits LP, Luikart G, Taberlet P. Estimating the probability of identity among genotypes in natural populations: cautions and guidelines. *Mol Ecol*. 2001;10:249–56
11. Bekele FL, Bidaisee GG, Allegre M. *et al.* Genome-wide association studies and genomic selection assays made in a large sample of cacao (*Theobroma cacao* L.) germplasm reveal significant marker-trait associations and good predictive value for improving yield potential. *PLoS One*. 2022;17:e0260907
12. Cull A, Joly DL. Development and validation of a minimal SNP genotyping panel for the differentiation of *Cannabis sativa* cultivars. *BMC Genomics*. 2025;26:83
13. Chen J, Zavala C, Ortega N. *et al.* The development of quality control genotyping approaches: a case study using elite maize lines. *PLoS One*. 2016;11:e0157236
14. Semagn K, Beyene Y, Makumbi D. *et al.* Quality control genotyping for assessment of genetic identity and purity in diverse tropical maize inbred lines. *Theor Appl Genet*. 2012;125:1487–501
15. Melo ATO, Bartaula R, Hale I. GBS-SNP-CROP: a reference-optional pipeline for SNP discovery and plant germplasm characterization using variable length, paired-end genotyping-by-sequencing data. *BMC Bioinformatics*. 2016;17:29
16. Salgotra RK, Stewart CN. Functional markers for precision plant breeding. *Int J Mol Sci*. 2020;21:4792
17. Nieves-Orduña HE, Müller M, Krutovsky KV. *et al.* Genotyping of cacao (*Theobroma cacao* L.) germplasm resources with SNP markers linked to agronomic traits reveals signs of selection. *Tree Genet Genomes*. 2024;20:13
18. Gouda AC, Warburton ML, Djedatin GL. *et al.* Development and validation of diagnostic SNP markers for quality control genotyping in a collection of four rice (*Oryza*) species. *Sci Rep*. 2021;11:18617
19. Ramirez-Ramirez AR, Bidot-Martínez I, Mirzaei K. *et al.* Comparing the performances of SSR and SNP markers for population analysis in *Theobroma cacao* L., as alternative approach to validate a new ddRADseq protocol for cacao genotyping. *PLoS One*. 2024;19:e0304753
20. Zhang D, Mischke S, Goenaga R. *et al.* Accuracy and reliability of high-throughput microsatellite genotyping for cacao clone identification. *Crop Sci*. 2006;46:2084–92
21. Osorio-Guarín JA. *et al.* Assessing genetic redundancy and diversity in Colombian cacao germplasm banks using SNP fingerprinting. *Front Plant Sci*. 2025;16:1632888
22. Steadman EM, Whitehouse KJ, Norton SL. Molecular characterization of *Brassica* genebank germplasm confirms taxonomic identity and reveals low levels and source of taxonomic errors. *Plant Genet Resour Charact Util*. 2022;20:194–202
23. Fang W, Meinhardt LW, Mischke S. *et al.* Accurate determination of genetic identity for a single cacao bean, using molecular markers with a nanofluidic system, ensures cocoa authentication. *J Agric Food Chem*. 2014;62:481–7

Downloaded from https://academic.oup.com/hr/article/13/7/uhag106/8528763 by guest on 03 August 2026

24. Hampe J, Schreiber S, Krawczak M. Entropy-based SNP selection for genetic association studies. *Human Genetics*. 2003;114:36–43
25. Brunel H, Gallardo-Chacón JJ, Buil A. *et al.* MISS: a non-linear methodology based on mutual information for genetic association studies in both population and sib-pairs analysis. *Bioinformatics*. 2010;26:1811–8
26. Gong L, Gao D, Zhang X. *et al.* REL-NPMI: exploring genotype and phenotype relationship of pancreatitis based on improved normalized point-by-point mutual information. *Comput Biol Med*. 2023;158:106868
27. Livingstone D, Royaert S, Stack C. *et al.* Making a chocolate chip: development and evaluation of a 6K SNP array for *Theobroma cacao*. *DNA Res*. 2015;22:279–91
28. Baek I, Cha M, Lim S. *et al.* Integrated phenotypic analysis, predictive modeling, and identification of novel trait-associated loci in a diverse *Theobroma cacao* collection. *BMC Plant Biol*. 2025;25:1050
29. Menne MJ, Durre I, Vose RS. *et al.* An overview of the global historical climatology network-daily database. *J Atmospheric Ocean Technol*. 2012;29:897–910
30. Pedregosa F, Varoquaux G, Gramfort A. *et al.* Scikit-learn: machine learning in python. *J Mach Learn Res*. 2011;12:2825–30
31. Shannon CE. A mathematical theory of communication. *Bell Syst Tech J*. 1948;27:379–423
32. Henning JA, Coggins J, Peterson M. Simple SNP-based minimal marker genotyping for *Humulus lupulus* L. identification and variety validation. *BMC Res Notes*. 2015;8:542
33. Kruskal JB. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*. 1964;29:1–27
34. Motamayor JC, Lachenaud P, da Silva e Mota JW. *et al.* Geographic and genetic population differentiation of the Amazonian chocolate tree (*Theobroma cacao* L.). *PLoS One*. 2008;3:e3311

Downloaded from https://academic.oup.com/hr/article/13/7/uhag106/8528763 by guest on 03 August 2026

Published by Oxford University Press 2026 on behalf of Nanjing Agricultural University. This work is written by (a) US Government employee(s) and is in the public domain in the US.
Horticulture Research, 2026, 13, 1–11
https://doi.org/10.1093/hr/uhag106
Article