

The importance of allele dosage: a simulation study on genomic prediction and response to selection in triploid hybrid breeding

Kajal Gupta

kgupta23@ncsu.edu

North Carolina State University

Kashish Grover

North Carolina State University

Hercules Pereira

North Carolina State University

Roberto Fritsche-Neto

North Carolina State University

Research Article

Keywords:

Posted Date: October 27th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11157292/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

The importance of allele dosage: a simulation study on genomic prediction and response to selection in triploid hybrid breeding

Kajal Gupta¹, Kashish Grover¹, Hercules Pereira¹, Roberto Fritsche-Neto¹

¹Department of Horticultural Science, North Carolina State University, Raleigh, North Carolina, United States

Abstract

Genomic prediction in polyploid crops requires decisions about how allele dosage is encoded at biallelic SNPs. In triploid breeding systems produced from tetraploid $\times$ diploid crosses- the mating design used commercially for seedless watermelon and other economically important crops- each locus can carry the alternative allele in zero, one, two, or three copies. This four-state genotype is frequently compressed to a diploid-like three-state encoding for compatibility with standard prediction pipelines. The long-term cost of this compression, and the conditions under which it becomes practically important, remain unquantified. We used stochastic simulation to isolate the effect of dosage encoding on prediction accuracy and long-term selection response in a triploid hybrid breeding program. We compared two encodings: the proper four-state dosage encoding (C: 0/1/2/3) and a diploidized three-state encoding that collapses the two heterozygous classes (B: 0/1/1/2). We evaluated these encodings across a 72-scenario factorial grid varying dominance, heritability, and training size. Overall, encoding C outperformed encoding B under the additive-plus-dominance kernel model in 66 of 72 scenarios. The difference was largest when the dominance proportion was low and shrank monotonically as dominance became more important. In a subsequent recurrent selection study, overall, encoding C produced 7.6% higher cumulative hybrid genetic value than encoding B at cycle 10 (30 years of breeding). However, the per-cycle prediction accuracy advantage of dosage-aware encoding is modest (0.028 accuracy units). Our results place a lower bound on the value of dosage-aware genotyping in polyploid breeding and identify the trait architectures for which the investment is most warranted.

Introduction

Genomic prediction has transformed quantitative breeding by enabling selection decisions based on genotype information before or in place of full phenotypic evaluation. Since the foundational demonstration by Meuwissen et al. (2001) that dense marker maps can predict breeding values with useful accuracy, the approach has been implemented across dairy cattle (Hayes et al. 2009), poultry, and a growing set of plant species (Crossa et al. 2017). The core statistical framework is well established for diploid organisms: mixed-model or ridge-regression prediction from a genotype matrix in which each individual's genotype at each biallelic locus is represented as the count of the alternative allele, taking values 0, 1, or 2. Kinship-based prediction using the additive genomic relationship matrix of VanRaden (2008) has become the default, with straightforward extensions to dominance (Vitezica et al. 2013; Muñoz et al. 2014) and epistasis (Jiang and Reif 2015) effects.

The extension of genomic prediction to polyploid crops has progressed more slowly and unevenly. Most polyploids of agricultural importance are autopolyploids or segmental allopolyploids in which each locus carries more than two allele copies per individual. In an autotetraploid, for example, a biallelic locus can carry the alternative allele in 0, 1, 2, 3, or 4 copies (dosages 0 through 4); a triploid similarly supports dosages 0 through 3, and an autohexaploid 0 through 6 (Endelman 2023). Standard diploid genomic prediction pipelines cannot directly represent this information because their genotype encodings assume only three possible states per locus (Endelman et al. 2018). Two general responses have emerged in the polyploid GP literature.

The first response is to develop dosage-aware prediction methods that retain the full multi-state genotype. VanRaden's additive relationship matrix generalizes naturally to arbitrary ploidy m by dividing the centering and scaling terms by m rather than by 2 (Endelman et al. 2018). Dominance kernels have been extended to polyploids using indicator functions for heterozygous states (Batista et al. 2022) or Fourier-based decompositions that partition digenic, trigenic, and tetragenic dominance in autotetraploids (Endelman 2023). These methods have been applied to autotetraploid potato (Endelman et al. 2018; Sverrisdóttir et al. 2017; Bourke et al. 2018), autotetraploid blueberry (Amadeu et al. 2020; Ferrão et al. 2021), autotetraploid alfalfa (Jia et al. 2021), and other systems. The general finding is that dosage-aware prediction yields modestly higher accuracy than encodings that discard dosage information, though the magnitude of the gain varies across species, traits, and population structures.

The second response is to sidestep the polyploid encoding problem by compressing dosage information into a diploid-compatible three-state encoding. This can take several forms. One common approach reads dosage from a genotyping platform designed for diploids and reports only three states (homozygous reference, heterozygous, homozygous alternative), effectively collapsing all intermediate dosages into the heterozygous class (Bourke et al. 2018). Another approach applies a downstream mapping to reduce dosage vectors: for autotetraploids, dosages 1, 2, and 3 might all map to a single heterozygous class; for triploids, dosages 1 and 2 might be similarly collapsed. These simplifications remain widely used in practice for several reasons (de Bem Oliveira et al. 2019; Amadeu et al. 2020). Dosage-aware calling requires sequencing depth sufficient to reliably distinguish adjacent dosage classes, which in turn requires either high-coverage sequencing or specialized calling algorithms that model the read-count distribution at each locus. Depending on platform and target coverage, per-sample genotyping costs for dosage-aware calling can be higher than for diploid calling of the same samples (Uitdewilligen et al. 2013) and requires more computationally intensive calling algorithms that model the read-count distribution at each locus (Voorrips et al. 2011; Serang et al. 2012; Gerard et al. 2018). The additional steps required to call and validate dosages typically extend the turnaround time from raw reads to a usable genotype matrix by days to weeks compared to diploid calling. For programs on a fixed genotyping budget, this cost-and-time premium translates directly into fewer samples genotyped per cycle and slower selection decisions, which are real constraints on program throughput.

The question of when dosage-aware encoding is worth its incremental cost and when it is not has not been systematically addressed. Published comparisons typically report accuracy differences from single empirical datasets in a specific species and trait context (Endelman et al. 2018; Sverrisdóttir et al. 2017; Amadeu et al. 2020; Ferrão et al. 2021; de Bem Oliveira et al. 2019), providing point estimates rather than a general characterization. Because prediction accuracy depends on multiple interacting factors: trait architecture, dominance structure, heritability, training population size, marker density, family structure, a single point estimate from one dataset cannot support general conclusions about when dosage-aware encoding matters most. Moreover, most empirical comparisons report only single-cycle prediction accuracy, whereas a prediction method's practical value to a breeding program depends on cumulative selection response over many cycles. A small per-cycle advantage that compounds over cycles can be operationally decisive; a larger per-cycle advantage that fails to compound might not justify a pipeline change.

Triploid hybrid breeding provides a particularly clean setting in which to address these questions. Triploid crops produced from crosses between tetraploid females and diploid males, including seedless watermelon (*Citrullus lanatus*) (Kihara, 1951; Rhodes & Zhang, 2008), triploid banana, sugar beet, and several ornamental species which represent a substantial commercial sector. In these programs, the triploid is the commercial product. Prediction of the triploid's genetic merit is therefore the object of selection, and the encoding of triploid genotypes is a decision every triploid breeding program must make. Triploids offer four dosage states (0, 1, 2, 3) and the distinction between the two triploid heterozygous classes (AAB, dosage 1; and ABB, dosage 2) is the specific information that any three-state diploidized encoding discards. Studying the cost of this compression in triploids, therefore, isolates the encoding effect in the simplest interesting case.

Two properties of the problem make it well suited to controlled simulation rather than purely empirical study. First, dosage encoding is a computational transformation of a known genotype rather than a biological process. Comparing encodings requires no biological replication, only a defined genotype-to-phenotype architecture and a way to apply each encoding to the same simulated individuals. Second, the effect of encoding is expected to depend on trait architecture and program design. Traits with substantial dominance may differ from purely additive traits; large training sets may reveal encoding advantages that small ones miss; and higher heritability may either amplify or attenuate the encoding gap. A factorial simulation across these dimensions can characterize the encoding-by-architecture interaction that no empirical dataset can resolve in isolation.

In this context, this study aimed to answer three questions. First, how large is the per-cycle prediction accuracy cost of dosage compression, and how does this cost depend on trait architecture and training design? Second, does the per-cycle accuracy difference translate into meaningful differences in long-run genetic gain, and if so, how does the encoding advantage propagate through recurrent selection? Third, under what conditions is dosage-aware encoding most valuable, so that programs facing a choice between dosage-aware and diploidized calling can prioritize their investment appropriately? Our results provided quantitative answers to each

question and placed a lower bound on the practical value of dosage-aware genotyping for polyploid breeding programs pursuing quantitative traits across multi-year horizons.

Materials and Methods

Overview

We used stochastic simulations to quantify the effects of triploid dosage encoding on genomic prediction accuracy and long-run genetic gain under recurrent selection. The study consisted of two experiments built on a shared simulation foundation (Figure 1). This foundation constructed a diploid genetic history, established a set of elite diploid parents, generated tetraploid founders by chromosome doubling, and locked a common genome and trait architecture across ploidies.

Experiment 1: evaluated genomic selection accuracy across a 72-scenario factorial sensitivity analysis, with 50 replicates per scenario, comparing full four-state (C) and diploidized three-state (B) dosage encodings under additive and additive-plus-dominance models.

Experiment 2: evaluated the two encodings in a 10-cycle recurrent selection scheme with 100 paired replicates, using identical starting populations and stochastic realizations to isolate the effect of dosage encoding on cumulative genetic gain.

All simulations were conducted in R (v4.5.1; R Core Team 2025), with AlphaSimR (v2.1.0; Gaynor et al. 2021) used for population genetics and meiosis, rrBLUP (v4.6.3; Endelman 2011) for variance component estimation. Figures were produced using ggplot2 (Wickham, 2016). Full simulation code and analysis scripts are available in the <https://github.com/Kajal0608/Simulation-Genomic-Prediction-on-Polyploids>

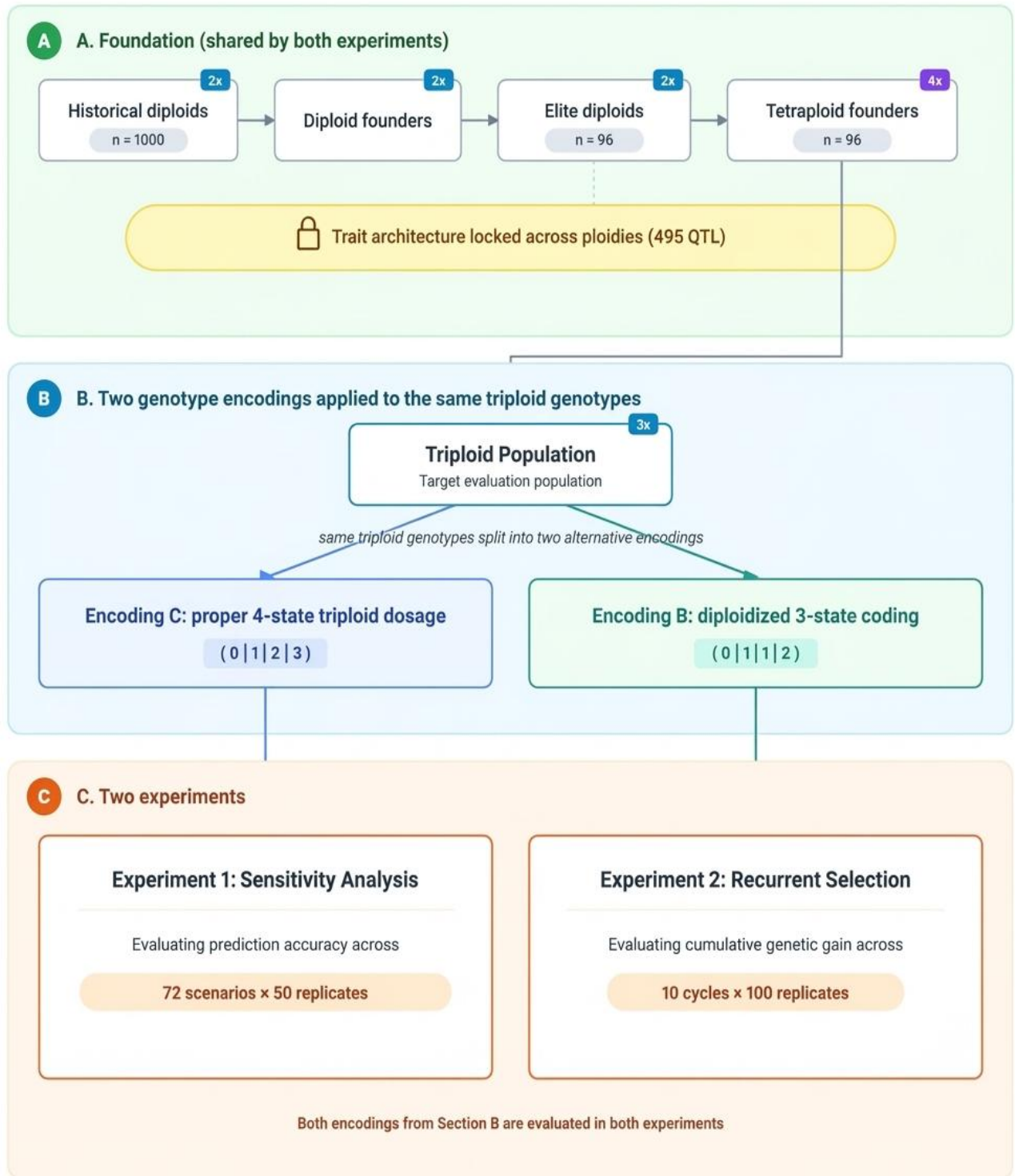


Figure 1: Study design comparing proper triploid dosage encoding (C) and diploidized encoding (B) in genomic prediction accuracy and recurrent selection response to selection.

Genome and marker map

The simulated genome was structured to reflect the chromosome number of *Citrullus lanatus* while remaining broadly relevant to polyploid crop breeding. Thus, we simulated 11 chromosomes, each carrying 318 segregating sites, for a total of 3,498 segregating loci. Each chromosome was represented on a genetic map of 100 centimorgans with uniformly spaced sites. Historical diploid haplotypes for 1,000 individuals were generated using the quickHaplo function of AlphaSimR, which produces an ancestral haplotype pool by drawing allele states independently at each site from a distribution parameterized to yield a specified minor allele frequency distribution. This approach substitutes computational speed for the more explicit coalescent-based simulation available through the runMacs function; the difference is immaterial for our purposes because the study's inference does not depend on the specific structure of long-range linkage disequilibrium in the founder pool.

A subset of 495 sites (45 per chromosome) was designated as quantitative trait loci (QTL). The remaining 3,003 sites (273 per chromosome) constituted a SNP genotyping chip. QTL sites and chip sites were mutually exclusive, so that the prediction problem was one of predicting genetic value at unobserved QTL from observed non-QTL marker information—a setting analogous to real breeding programs in which causal variants are not directly measured. The SNP chip was added to both the diploid and tetraploid simulation parameter objects via addSnpChip function, using the same site positions in both.

Founder populations

We constructed the diploid founder population as follows. From the 1,000 historical diploid haplotypes, we created a diploid simulation parameter object (SP2) with SimParam\$new. From this SimParam object, we generated a diploid founder population (firstPop) of 1,000 individuals via the newPop function. We phenotyped this population once under a fixed-trait architecture, computed each individual's phenotypic value, and selected the top 96 individuals based on their phenotypic values. This selection step represents the elite diploid germplasm from which tetraploid founders would realistically be constructed in a commercial breeding program: tetraploid parents are not typically induced from unselected base populations, but from lines that have already demonstrated agronomic value at the diploid level. The number 96 was chosen to match a two-tray tissue-culture batch size in typical commercial practice and to provide sufficient tetraploid founder diversity for downstream selection cycles.

Tetraploid founders were produced by doubling the chromosomes of the 96 elite diploid parents using the doubleGenome function. The resulting tetraploid individuals retained the diploid genetic map but carried four copies of each chromosome, with each pair of chromosomes descended from the corresponding pair in the diploid ancestor. Tetraploid haplotypes were extracted from the doubled population using the pullSegSiteHaplo function and reassembled into a new MapPop object retaining the original genetic map. A tetraploid simulation parameter object (SP4) was then created from this MapPop.

The trait architecture from SP2 was copied into SP4 (so that additive QTL effects were taken directly without rescaling), and environmental variance was set identically in both SimParam objects via setVarE. This construction ensured that QTL positions, additive effect sizes, dominance degrees, and environmental variance were identical across ploidies. Any observed difference in prediction accuracy or genetic gain between the two encodings could therefore be attributed to the encoding transformation itself, not to any unaccounted difference in the underlying genetic architecture.

The tetraploid quadrivalent pairing probability was set to 0.3. We chose this as a moderate modeling assumption to allow a mixture of bivalent and quadrivalent pairing at meiosis (Sybenga 1975; Bourke et al. 2015)., appropriate to segmental autotetraploids. This value was fixed across all downstream simulations.

Trait architecture

Each simulated trait had an additive-plus-dominance architecture with 495 QTL as specified above. Three architectural parameters were varied across scenarios in Experiment 1 and fixed at a baseline setting in Experiment 2.

Additive QTL effects were sampled from a gamma distribution to represent a genetic architecture with many small-effect loci and a few large-effect loci. The effects were then rescaled to set the founder-population additive variance to 1. This scaling provided a common reference for specifying environmental variance at each target heritability without additional renormalization across scenarios.

Dominance was parameterized by two quantities: the mean dominance degree, defined as the ratio of the dominance effect to the additive effect averaged across QTL; and the variance of the dominance degree across QTL which controls the spread of dominance degrees around the specified mean. Dominance effects at each QTL were drawn from a normal distribution with mean equal to the mean dominance degree multiplied by the additive effect at that QTL and standard deviation equal to the square root of the dominance-degree variance multiplied by the additive effect.

Environmental variance (V_E) was derived from a target base heritability h^2_{base} according to:

$$V_E = V_A \times (1 - h^2_{base}) / h^2_{base}$$

with $V_A = 1$ based on the additive-effect scaling described above. This V_E value was applied identically to both SP2 and SP4. Because V_E was fixed across ploidy levels rather than heritability being fixed, the realized heritability could differ modestly between diploid and tetraploid (and subsequently triploid) populations, reflecting the fact that in real breeding programs, environmental variance is a property of the phenotyping conditions rather than of the ploidy of the material being evaluated.

Triploid hybrids

Triploid individuals were produced by crossing tetraploid females with diploid males. This combines diploid gametes from the tetraploid parent and haploid gametes from the diploid parent to produce triploid offspring. The resulting triploid population objects carry ploidy $m = 3$ and dosage vectors taking values 0, 1, 2, or 3 at each locus.

For Experiment 1, we generated triploid populations by crossing tetraploid and diploid parents. The number of crosses and progeny per cross were varied across scenarios (see below). For Experiment 2, we generated training and evaluation triploid populations separately at each cycle from the current-cycle tetraploid and diploid parent pools.

Genotype encodings

The two encodings compared throughout the study are defined as follows.

Encoding C (complete dosage). For each of the 3,003 chip SNPs, the triploid individual's dosage is recorded as an integer in $\{0, 1, 2, 3\}$, corresponding to the count of the alternative allele across the three homologous chromosome copies. The dosage matrix has dimensions $n \times 3,003$ and takes values in $\{0, 1, 2, 3\}$. The ploidy parameter m used in downstream kernel construction is 3.

Encoding B (diploidized). The four-state dosage vector is compressed to a three-state vector by mapping $0 \rightarrow 0$, $1 \rightarrow 1$, $2 \rightarrow 1$, and $3 \rightarrow 2$. Under this mapping, the two heterozygous triploid classes, AAB (dosage 1) and ABB (dosage 2), collapse into a single heterozygous category. The dosage matrix has dimensions $n \times 3,003$ and takes values in $\{0, 1, 2\}$. The ploidy parameter m used in downstream kernel construction is 2.

Both encodings were applied to the same triploid individuals with the same underlying genotypes. The B encoding therefore represents the information loss that occurs when a diploid-oriented genotype-calling pipeline is applied to triploid genotypes. Because both encodings start from the same perfectly known dosages, the comparison isolates the information content of the encoding itself, independent of any measurement error that a real dosage-calling pipeline might introduce.

Kernel construction and prediction models

For each encoding, we built two matrices describing genetic similarity between individuals: an additive kernel and a dominance kernel.

The additive kernel measures how similar individuals are in the alleles they carry. It follows the standard formulae of VanRaden (2008), generalized to arbitrary ploidy by Endelman et al. (2018):

$$K_A = \frac{\mathbf{W}\mathbf{W}^T}{m \sum_j p_j (1 - p_j)}$$

where the matrix $\mathbf{W}$ contains the individual dosages centered on the population mean allele frequency at each SNP, and m is the ploidy of the encoding, three for encoding C and two for

encoding B. Because the centering and the scaling both depend on ploidy, the two encodings produce kernels on the same scale which allows their prediction accuracies to be compared directly.

The dominance kernel measures how similar individuals are in their heterozygosity that is how many loci where they carry both allele types. For encoding C, we counted a locus as heterozygous if the triploid carried the alternative allele in one or two copies (dosages 1 or 2). For encoding B, we counted a locus as heterozygous if the individual was coded as 1 in the diploidized encoding. In each case, the heterozygosity indicator was centered on its expected value under Hardy-Weinberg equilibrium, and the resulting matrix was scaled to yield the dominance kernel.

We then fit two model variants for each encoding. The additive-only model used the additive kernel alone. The additive-plus-dominance model used the sum of the additive and dominance kernels. Both models were fit using the rrBLUP package (Endelman 2011), which estimates the ratio of environmental to genetic variance from the training population and uses it to predict genetic values in a held-out validation set.

Prediction accuracy was measured as the Pearson correlation between predicted and true simulated genetic values in the validation individuals. Because we know the true genetic values in a simulation, this measure directly captures how well each model recovers the underlying signal, without the noise that would arise when validating against phenotypes in real data.

Experiment 1: Factorial sensitivity of prediction accuracy. Measured how the encoding-related accuracy gap depends on trait architecture and training design. The factorial grid comprised five factors, totaling 72 scenarios replicated over 50 times:

Mean dominance degree: {0.2, 0.6, 1.0} (three levels)

Variance of dominance degree: {0.1, 0.5} (two levels)

Base heritability: {0.05, 0.12, 0.30} (three levels)

Number of crosses: {150, 400} (two levels)

Progeny per cross: {2, 5} (two levels)

Within each scenario, we followed the procedure below. We first regenerated the founder pool using a random seed. From this pool, we produced a training population of triploid individuals by crossing tetraploid mothers with diploid fathers randomly drawn from the elite parent set. We varied the number of crosses and the number of progenies per cross across scenarios (as described above), yielding training populations of varying sizes. Each triploid was assigned a phenotype using the environmental variance set for that scenario.

Prediction accuracy was then estimated by five-fold cross-validation grouped by cross so that no full-sib family appeared in both the training and the validation partitions of any fold. Within each fold, we fitted both prediction models (additive-only and additive-plus-dominance) under both

encodings (complete and diploidized), and predicted the genetic values of the held-out validation individuals. Prediction accuracy for a given scenario, model, and encoding was the average correlation across the five folds.

The main outcomes were the four prediction accuracies obtained from the model-encoding combinations. We also calculated the encoding difference for each model as the difference in accuracy between the proper triploid encoding and the diploidized encoding. This was evaluated for both the additive-only model and the additive-plus-dominance model. Moreover, for each scenario, we estimated the realized additive and dominance variances from the simulated genetic values. We also calculated the dominance proportion as the fraction of total genetic variance explained by dominance. Finally, we fit exploratory linear models to test how the four factorial factors and dominance proportion affected the encoding difference under the additive-plus-dominance model.

Experiment 2: Recurrent selection. Measured the cumulative effect of dosage encoding on genetic gain across a 10-cycle recurrent selection program, representing 30 years of breeding. Each replicate consisted of a 10-cycle run under both encodings, using paired seeds so that both encodings started from the same founder pool and experienced the same stochastic realizations of mating and selection, differing only in marker-dosage encoding during model training and prediction. The experiment comprised 100 replicates, a total of 2,000 breeding-cycle simulations with replicate-level seeds spaced sufficiently to avoid overlap.

Each selection cycle involved three steps.

We first generated a new training population for genomic prediction. The population consisted of 1,500 triploid individuals produced by crossing 300 tetraploid mothers with diploid fathers selected at random from the current parent pools. Five progenies were generated per cross. The triploids were genotyped using the SNP chip and evaluated under both the complete and diploidized encodings. They were then phenotyped under the trait architecture specified for the experiment and used to fit an additive-plus-dominance prediction model.

Second, we predicted the genetic value that each parent would contribute to future triploid offspring. Direct estimation would require simulating and genotyping many test-cross progeny for each parent. This approach was computationally infeasible for 800 parents across 2,000 breeding-cycle simulations. Instead, we used the expected SNP dosage of triploid offspring from a tetraploid $\times$ diploid cross. The expected dosage was calculated as two thirds of the tetraploid parent's dosage plus one third of the average dosage in the diploid parent pool. This expected dosage vector was then passed through the fitted kernel model to predict the genetic value of each parent. We used this predicted value as the parent's general combining ability (GCA) for selection. This shortcut was used only for parent evaluation. The prediction model was still trained on simulated triploid genotypes and phenotypes at every cycle. Thus, the breeding pipeline retained a full triploid training step while substantially reducing computational requirements.

Third, we selected parents and generated the next breeding cycle. Within each pool, the 40 parents with the highest predicted GCA were selected and randomly crossed to produce the next generation of 400 tetraploids and 400 diploids. We also generated a separate evaluation population from 200 crosses with three progenies per cross. This resulted in 600 triploid individuals. The mean true genetic value of this population was recorded as the outcome for each breeding cycle.

Comparing scenarios

For Experiment 1, we reported the mean and standard deviation of the encoding gap across the scenarios, along with the count of scenarios favoring each encoding. We fit linear models of the encoding gap on the factorial predictors and the realized dominance proportion, and we present marginal-mean plots of the gap against each key predictor.

For Experiment 2, the primary outcome was the mean hybrid genetic value of the evaluation population at cycle 10. The two encodings were compared using a paired-difference approach. We reported the mean and standard deviation for each encoding and the paired difference ($C - B$) across replicates. A 95% confidence interval and paired t-test were used to assess the difference. The Wilcoxon signed-rank test was also performed as a nonparametric alternative. We calculated the proportion of replicates in which C had a higher genetic value than B at cycle 10 to assess the consistency of the encoding effect across breeding-program replicates. Cumulative gain relative to cycle 1 was recorded for each cycle under both encodings. Differences between encodings were also calculated for each cycle to assess how the encoding effect changed over the course of selection. All analyses were performed in R using base functions for the paired t-test and Wilcoxon signed-rank test (Wilcoxon, 1945).

Results

Prediction accuracy across the sensitivity grid

Encoding C outperformed encoding B in most scenarios under both prediction models. Under the additive-plus-dominance model, encoding C was more accurate in 66 of 72 scenarios (92%), with a mean advantage of 0.028 accuracy units (standard deviation 0.017; range -0.01 to 0.06). Under the additive-only model, encoding C was more accurate in 66 of 72 scenarios (92%), with a mean advantage of 0.028 (standard deviation 0.017; range -0.01 to 0.06). The two models produced nearly identical patterns, indicating that the advantage of encoding C is not driven by the dominance kernel alone.

Absolute prediction accuracies were substantial across the grid: 0.60 to 0.85 under encoding C (mean 0.74) and 0.58 to 0.82 under encoding B (mean 0.71). The average advantage of 0.028 units therefore represents a modest but consistent shift on top of a substantial baseline.

Encoding C showed higher prediction accuracy than encoding B across nearly all the 72 scenarios (Figure 2). Most points fell above the diagonal line of equal accuracy, indicating a consistent advantage for encoding C. The six scenarios in which encoding B was marginally more accurate

had differences ranging from -0.001 to -0.01 units. These scenarios occurred across different combinations of dominance degree, heritability, and training size, with no clear pattern.

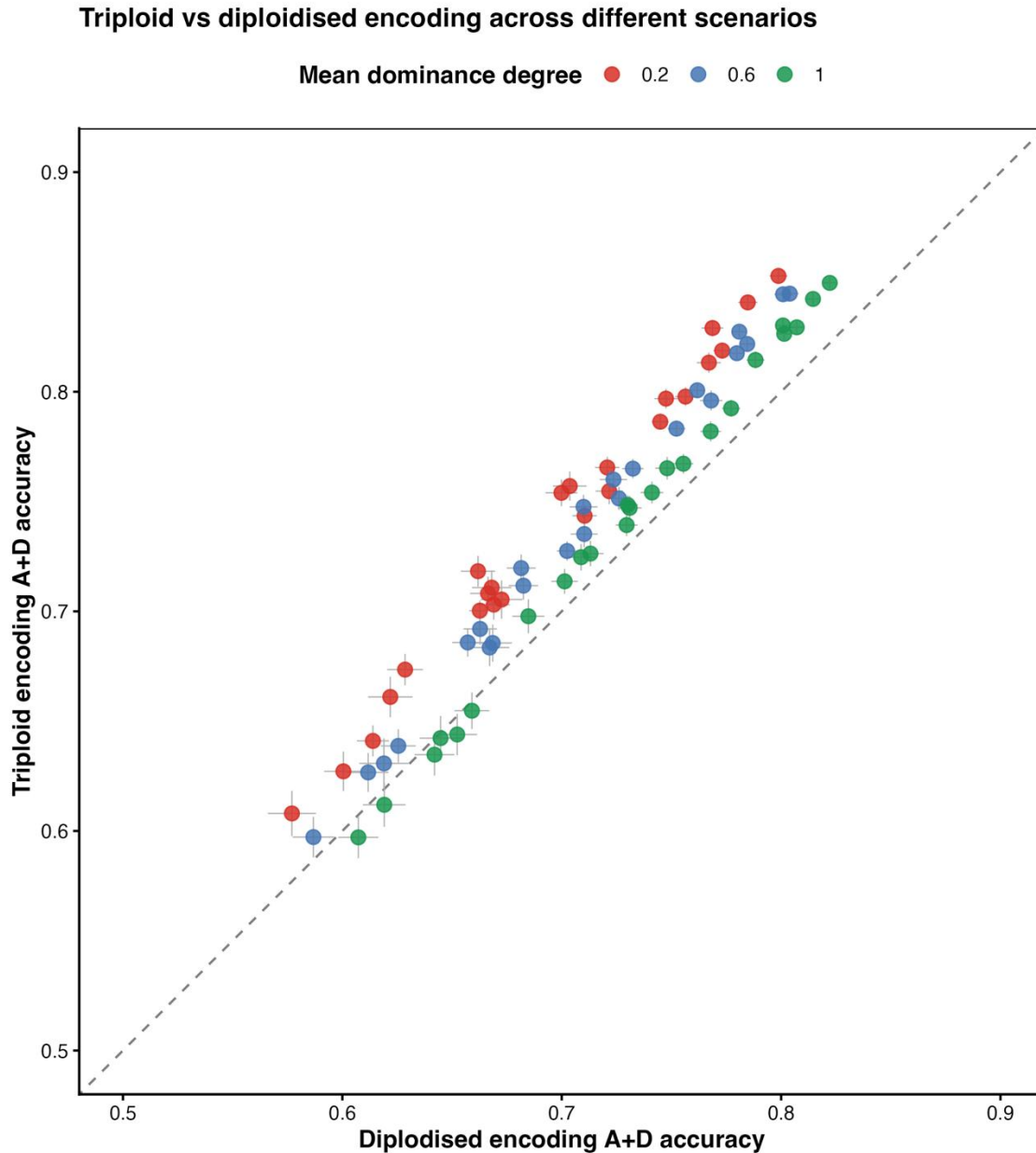


Figure 2: Triploid encoding outperforms diploidized in prediction accuracy across 66 of 72 sensitivity scenarios.

Effect of training population size, mean dominance degree, and heritability

The advantage of encoding C over encoding B varied systematically with three factors in the sensitivity grid: training population size, the trait's mean dominance degree, and baseline heritability (Figure 3A-C).

Prediction accuracy increased substantially with training population size under both encodings (Figure 3A). Under encoding C, accuracy increased from 0.64 with 300 training individuals to 0.83 with 2,000 individuals. Under encoding B, accuracy increased from 0.63 to 0.79 across the same range. The encoding C advantage increased from 0.01 to 0.04 as training population size increased. This represents a fourfold increase in the encoding advantage. A linear model including all factorial predictors showed a significant positive coefficient for the logarithm of training population size, confirming this trend after accounting for the other factors.

Mean dominance degree had an asymmetric effect on the two encodings (Figure 3B). Under encoding C, accuracy remained almost the same as dominance increased. Under encoding B, accuracy rose modestly in the opposite direction, from 0.70 at low dominance to 0.73 at high dominance. The two encodings therefore converged as dominance strengthened, and the encoding advantage shrank from 0.04 at low dominance to 0.01 at high dominance.

Baseline heritability had a small effect on both encodings (Figure 3C). Accuracy decreased as heritability increased under both encodings. Under encoding C, accuracy dropped from 0.75 at heritability 0.05 to 0.72 at heritability 0.30. Under encoding B, it dropped from 0.72 to 0.69 across the same range. The gap between the two encodings stayed at around 0.03 units across all heritability levels.

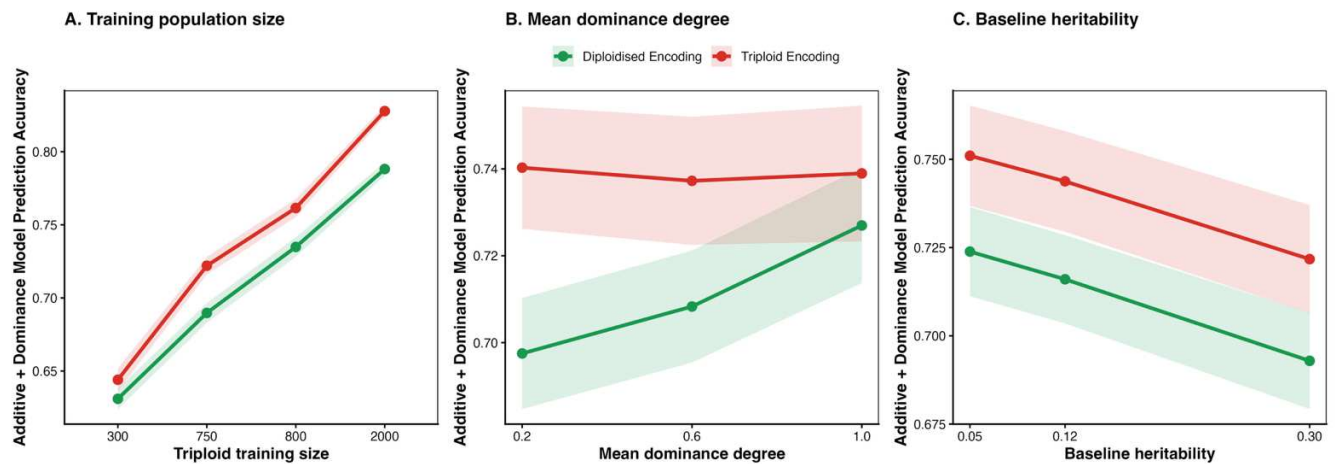


Figure 3: The encoding gap grows with training size (3A), is largest at low mean dominance degree (3B), and persists across heritability levels (3C).

Encoding gap drivers

The advantage of encoding C over encoding B decreased steadily as dominance contributed a larger share of total genetic variance (Figure 4). Averaging across scenarios, the advantage fell from about 0.05 units when dominance accounted for only 5% of genetic variance to near zero with a few scenarios slightly negative, when dominance accounted for 65%. The trend was consistent across all three heritability levels indicating that this pattern is not confounded with

heritability. All six scenarios in which encoding B was marginally more accurate than encoding C occurred at dominance proportions above 0.5.

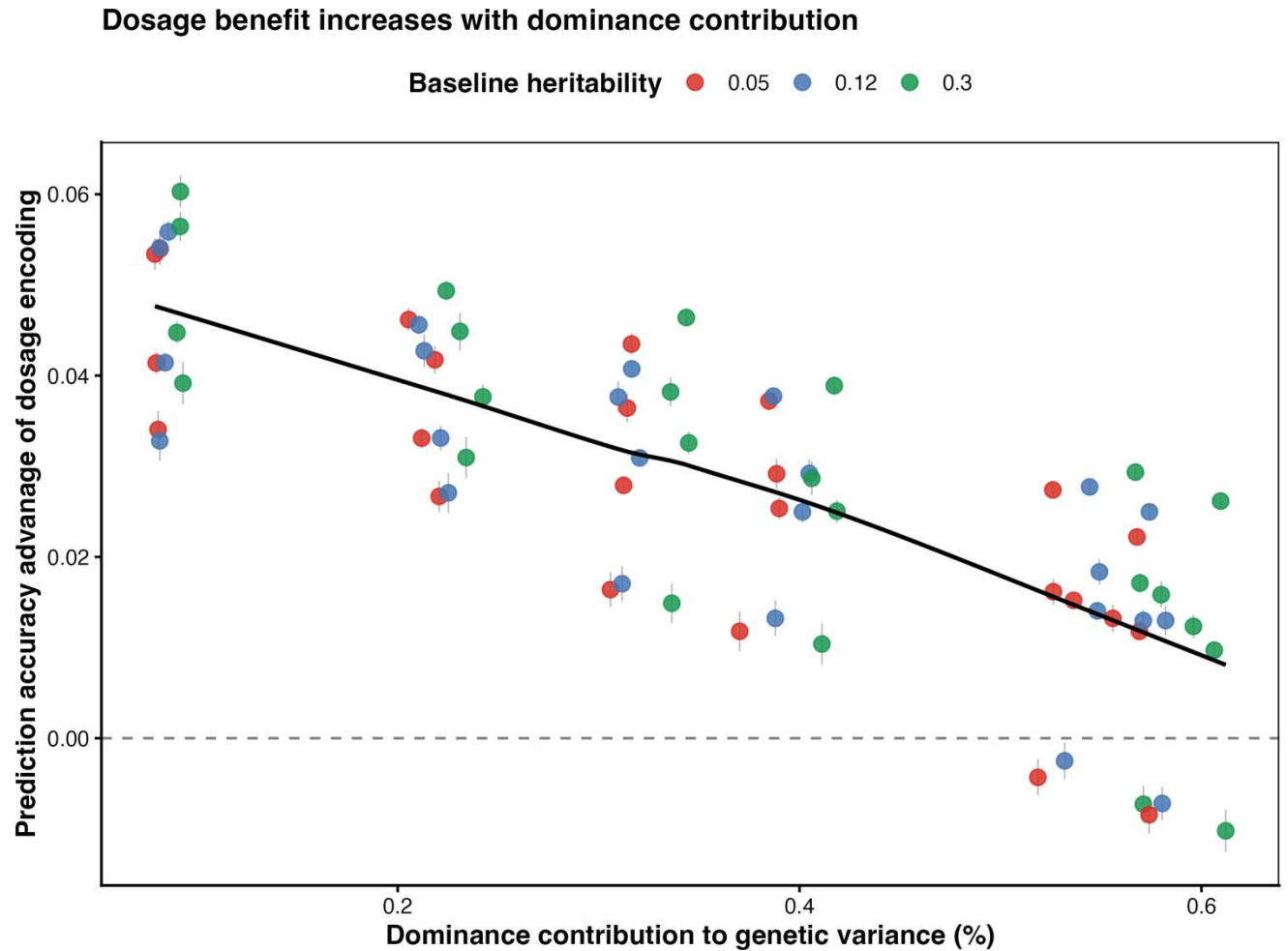


Figure 4: The encoding gap decreases monotonically as the dominance proportion of total genetic variance increases.

Long-term effects on breeding populations

Overall, encoding C produced greater cumulative genetic gain than encoding B (Figure 5). The paired design used the same founder populations, training populations, and random mating and selection events for both encodings within each replicate. Therefore, differences in genetic gain were attributable to the encoding choice.

The two encodings started with similar hybrid genetic values. Mean genetic value was 2.72 under encoding C and 2.56 under encoding B, giving a paired difference of 0.16. By cycle 10, the difference had increased substantially. Mean cumulative hybrid genetic value reached 8.67 under encoding C compared with 8.06 under encoding B. The mean paired difference was 0.61 (95% CI:

0.55 to 0.68). This corresponds to a 7.6% greater genetic value under encoding C relative to encoding B.

The difference at cycle 10 was statistically significant based on both tests. The paired t-test gave ($t = 18.93$), ($df = 99$), and ($p < 10^{-34}$). The Wilcoxon signed-rank test gave the same conclusion with ($V = 5041$) and ($p < 10^{-17}$).

Cumulative hybrid gain

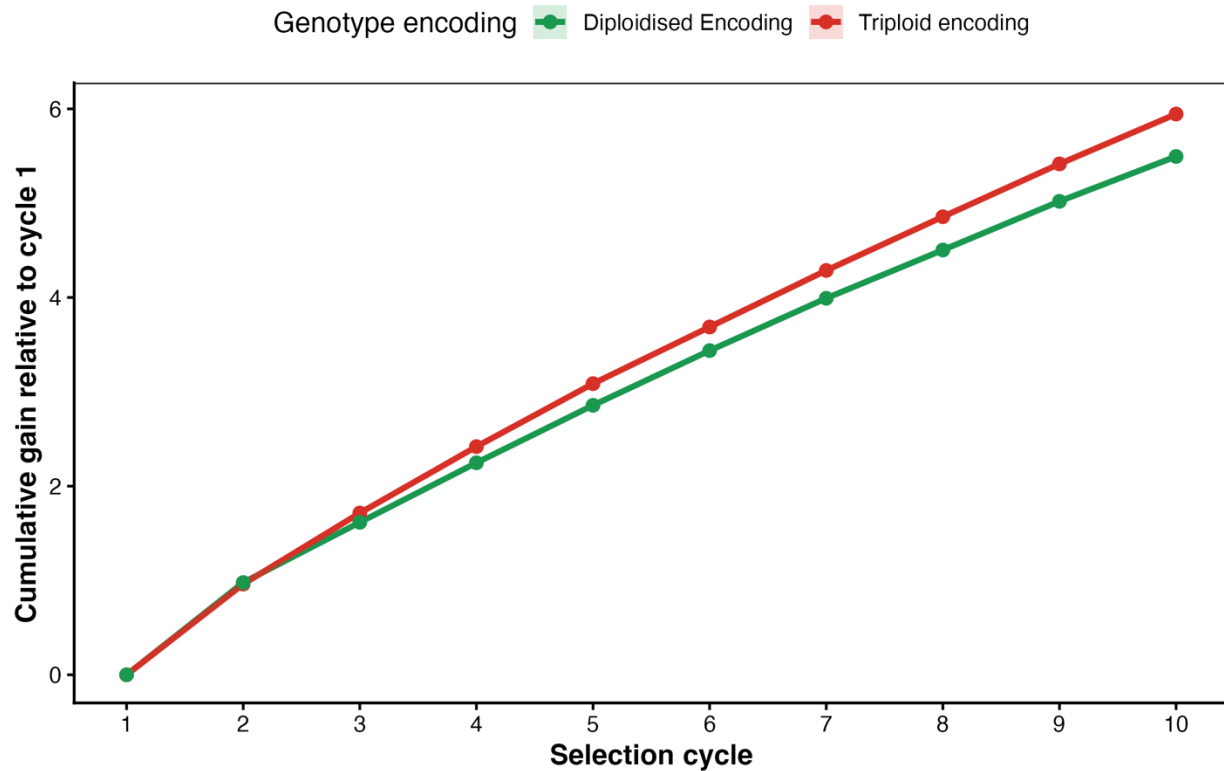


Figure 5: Cumulative hybrid genetic gain diverges progressively between encodings across 10 selection cycles.

The advantage of encoding C over encoding B held not just on average but in nearly every simulated breeding program. Across the 100 paired replicates, 98 produced higher hybrid genetic values under encoding C than under encoding B, with only two replicates showing similar outcomes for the two encodings (Figure 6).

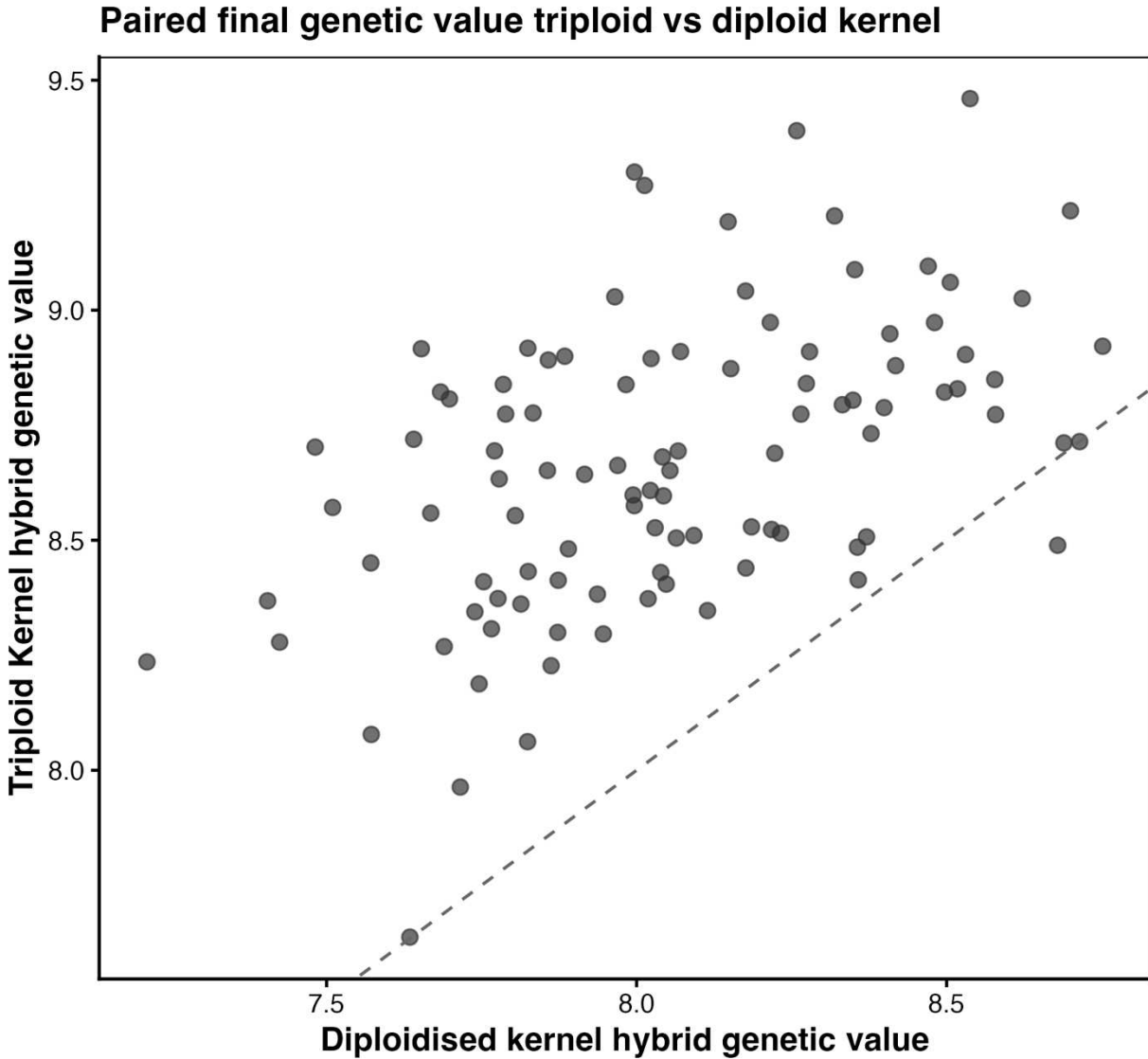


Figure 6: Encoding C outperforms encoding B in 98 of 100 paired recurrent selection replicates.

Discussion

A modest per-cycle accuracy difference compounds into considerable long-term genetic gain

This study addressed an unresolved question in polyploid genomic prediction: how much information is gained by using dosage-based encoding, and whether its value depends on trait architecture and selection duration. The two experiments provided complementary evidence.

The sensitivity analysis from Experiment 1 showed that dosage-aware encoding consistently produced higher prediction accuracy than diploidized encoding across the sensitivity grid. The mean advantage was modest and broadly consistent with previous studies in autotetraploid systems, which have reported accuracy improvements of a few percentage points from dosage-

aware encoding compared with diploidized approaches (Endelman et al. 2018; Sverrisdóttir et al. 2017; Amadeu et al. 2020; Ferrão et al. 2021). Our contribution is not the direction of the effect; this was expected from prior work, but rather the systematic characterization of when the advantage is large and when it is small across a broad range of trait architectures and training designs. An advantage of this size could reasonably be judged too small to justify the additional cost of dosage-aware genotyping when evaluated over a single selection cycle, as it is comparable to improvements from better phenotyping precision, higher marker density, or an alternative prediction model.

The recurrent selection experiment (Experiment 2) provided a different perspective on the same encoding effect. The small per-cycle advantage of encoding C accumulated across cycles into a substantial and statistically significant difference in genetic gain, occurring consistently across independent replicates. The magnitude of this compounding is worth understanding, because it is central to how our findings should be interpreted. The relationship between per-cycle accuracy and cumulative gain is not linear. Even a modest per-cycle advantage in prediction accuracy can amplify substantially when carried through a loop of recurrent selection. The mechanism is straightforward but easy to underestimate. In any single cycle, a small improvement in prediction accuracy translates into a slightly better ranking of candidate parents and, therefore, slightly better parent selection. The selected parents then contribute to the next generation of the breeding population, and the improvement in that next generation carries forward into the following selection cycle. In the following cycle, the improved population supports both a slightly higher mean genetic value in the training set and, in some cases, slightly better prediction accuracy when the training population itself is drawn from the improved pool. The cycle repeats. What began as a small per-cycle advantage in prediction accuracy becomes a widening gap in the mean genetic value of the breeding population, one cycle at a time, and the gap grows more quickly than a linear extrapolation of the per-cycle advantage would suggest.

This compounding performance is a general feature of selection response theory (Falconer and Mackay 1996), not a special property of dosage encoding. It has been demonstrated in genomic selection simulations for cattle (Gorjanc et al. 2018), for wheat (Bassi et al. 2016), and in two-part selection schemes for inbred crops (Gaynor et al. 2017). Its magnitude, however, has not previously been quantified for dosage-encoding choices in polyploid breeding. In our study, a per-cycle accuracy advantage of a few percentage points accumulated across ten cycles into a program-level advantage roughly twenty times larger, an amplification whose full magnitude only becomes apparent when the methodology is evaluated over the horizon on which breeding programs operate.

These two results together support what we regard as the study's primary contribution: the practical value of a genomic prediction methodology cannot be reliably estimated from single-cycle prediction accuracy comparisons alone. Cross-validation studies that report small accuracy advantages of one method over another in a specific dataset and then leave the reader to decide

whether the difference is worthwhile (Crossa et al. 2017), systematically underestimate the practical value of methods that produce small but consistent per-cycle improvements.

The dominance reversal and its mechanistic interpretation

The most subtle finding of the sensitivity analysis was that the encoding advantage depends on the realized dominance proportion of the trait. The encoding advantage was largest at low dominance proportions and decreased steadily as dominance contributed a larger share of the total genetic variance. This pattern reverses a natural first guess about when dosage encoding should matter most. If the comparison between the two encodings specifically concerns the distinction between the two heterozygous triploid classes - AAB and ABB - and if these two classes differ primarily in their dominance contribution, one might predict that the encoding advantage should grow with the importance of dominance to the trait. The data show the opposite.

In this context, we propose two complementary mechanisms for this pattern. Neither can be definitively confirmed from our current experiments, but both are consistent with the observed data. The first concerns what information the diploidized encoding preserves and what it discards. Encoding B does not eliminate all information about the two heterozygous triploid classes; it collapses AAB and ABB into a single heterozygous category but retains coarse information about the overall allele count across the genome. What it specifically discards is the *within-heterozygote additive contrast* - the fact that AAB carries one fewer copy of the alternative allele than ABB, and therefore contributes less to the additive breeding value. This within-heterozygote additive contrast is genetic information that becomes visible only when the full triploid dosage is retained. When most of a trait's genetic variance is additive, this contrast constitutes a proportionally larger fraction of the total signal, and encoding B's loss of it produces a proportionally larger accuracy penalty. When dominance dominates the genetic architecture, the within-heterozygote additive contrast is a smaller fraction of the total signal, and its loss matters correspondingly less. This interpretation is consistent with theoretical decompositions of polyploid genetic variance into additive and non-additive components (Endelman et al. 2018; Batista et al. 2022).

The second mechanism concerns the role of the dominance relationship matrix in the prediction model. This matrix is constructed from heterozygosity information, which is broadly equivalent between the two encodings - both encodings can identify heterozygous states at each locus, though with slightly different labeling. When dominance is strong, the dominance relationship matrix captures a large fraction of the total genetic signal in both encodings, and this contribution is largely encoding invariant. Encoding C's advantage over encoding B is therefore concentrated in the additive relationship component of the model, which contributes proportionally less to total accuracy when dominance is strong. This pattern of dominance kernels absorbing non-additive variance in a relatively encoding-invariant manner has been observed empirically in autotetraploid blueberry (Amadeu et al. 2020; Ferrão et al. 2021) and is consistent with the theoretical framework laid out by Vitezica et al. (2013) for dominance-augmented diploid prediction models.

The two mechanisms are complementary rather than competing, and both predict the direction and rough magnitude of the observed pattern. Distinguishing between them empirically would require additional experiments, for example, refitting encoding B under an additive-only model to isolate the additive component of the encoding advantage, or testing alternative diploidization schemes that discard different subsets of the dosage information.

The mechanistic reasoning we have offered is specific to the diploidization scheme we tested, in which the ABB class is collapsed onto the AAB class. Alternative diploidization schemes could produce qualitatively different dependence on dominance. Collapsing the ABB and BBB classes rather than the AAB and ABB classes, for instance, would discard different genetic information and might interact differently with the additive and dominance components of the trait. Programs that choose among practical diploidization strategies for triploid analysis should be aware that the encoding effect we characterize here is specific to one such strategy.

Implications beyond dosage encoding

For the compounding advantage to be meaningful for a real breeding program, the selection cycle must be translated into calendar time and evaluated under realistic operational constraints. A single selection cycle in a reciprocal recurrent selection program for triploid watermelon involves recombining both the tetraploid and diploid parent pools, producing the triploid training population from crosses between them, conducting field trials to phenotype the training triploids, genotyping, and finally selecting the next generation of parents. Each of these stages depends on watermelon's roughly 5-month generation time (Wehner 2008), on seasonal field-trial windows and nursery scheduling, and on weather gaps between stages. Under realistic program-management assumptions, one full selection cycle requires at least 3 years of program time, and the 10 selection cycles of Experiment 2 therefore represent approximately 30 years of breeding effort. This timeline is consistent with the recognized long cycle time in traditional triploid watermelon breeding (Kong et al. 2025). Over this horizon, encoding C produces approximately 7.6% more cumulative genetic gain than encoding B, provided the comparison is made cycle for cycle with equal numbers of genotyped individuals and equal cycle lengths.

This last condition of equal cycle length and equal per-cycle cost is the most important caveat to our results. Our simulation compares the two encodings under the assumption that both are equally cheap and equally fast, which is unlikely to hold in practice. Dosage-aware genotyping introduces two operational asymmetries that our simulation does not model. First, the turnaround from raw sequencing data to a usable dosage-called genotype matrix is typically longer than for diploid calling, days to weeks longer because dosage calling requires higher sequencing depth and more sophisticated bioinformatics (Uitdewilligen et al. 2013; Voorrips et al. 2011; Serang et al. 2012; Gerard et al. 2018). In a program on a fixed timeline, this extended turnaround could add meaningful time to each cycle and therefore reduce the annual rate of gain in absolute terms. Second, the per-sample cost of dosage-aware genotyping is typically higher than for diploid calling, depending on the sequencing platform and target coverage. In a program on a fixed budget, this cost premium translates directly into fewer individuals genotyped per cycle, which reduces both selection intensity and the effective genetic variability available for selection.

Combined through the breeder's equation, these two operational effects can substantially shrink the encoding advantage we observe under our idealized conditions and, under tight budget constraints or long turnaround times, could reverse the ranking entirely. A program that must choose between genotyping many more diploid-called individuals per cycle at low cost, versus fewer dosage-called individuals per cycle at higher cost and longer turnaround, may find that the higher throughput of the diploid pipeline produces a competitive or even superior outcome despite its lower per-cycle accuracy. Our simulation therefore quantifies the encoding effect under ideal conditions and should be interpreted as a ceiling estimate for the advantage of dosage-aware calling; realistic estimates for any specific program require folding local cycle length, budget, sample throughput, and selection intensity into the calculation.

Under favorable conditions where programs pursue quantitative traits with substantial additive genetic variance, and where the incremental cost and turnaround of dosage-aware genotyping can be absorbed by existing infrastructure, the annual rate-of-gain improvement is large enough to justify serious cost-benefit analysis. In such programs, particularly those producing multiple varieties per year with substantial associated seed sales, a compounding improvement in the rate of genetic gain across selection cycles can translate into value that readily exceeds the genotyping cost premium. In programs pursuing traits with weaker additive components or operating under tight budget constraints that would substantially reduce sample throughput, the case for dosage-aware genotyping is correspondingly weaker. The cost side of this calculation depends heavily on local infrastructure, personnel capacity, target traits, and budget. No single simulation study can address these variables. Our contribution is to provide the quantitative foundation for such program-specific decisions and to identify the trait architectures under which dosage-aware encoding is most likely to deliver a meaningful gain.

Limitations and future work

Our simulations assumed known allele dosages and focused only on genotype encoding. In practice, genotype calling errors associated with sequencing depth and calling algorithms may affect dosage accuracy (Serang et al. 2012; Gerard et al. 2018). Thus, our results may underestimate the potential advantage of dosage-aware encoding, although inaccurate dosage estimates could reduce this benefit. We also assumed equal genotyping costs and breeding-cycle lengths across strategies. In practice, dosage estimation may increase costs and processing time, potentially reducing the number of genotyped individuals by two- to five-fold and, consequently, genetic variability and selection intensity. Incorporating these factors into the breeder's equation could reduce or even reverse the observed advantage of dosage-aware encoding under budget constraints. Furthermore, we considered a single additive-plus-dominance architecture, fixed marker panel, tetraploid founders generated by chromosome doubling, and a simplified recurrent selection scheme. Alternative genetic architectures, marker densities, founder populations, and breeding strategies may change the magnitude of the differences but are unlikely to alter the conclusion that dosage-aware encoding is at least as effective as diploidized encoding under the conditions evaluated.

Future studies should evaluate alternative diploidization strategies and prediction approaches for identifying superior diploid parents before chromosome doubling, potentially reducing breeding costs while maintaining prediction accuracy. More broadly, genomic prediction methods should

be evaluated across multiple recurrent selection cycles, as small differences in prediction performance can accumulate into substantial long-term genetic gains.

References

- Amadeu, R. R., Ferrão, L. F. V., Oliveira, I. D. B., Benevenuto, J., Endelman, J. B., & Munoz, P. R. (2020). Impact of dominance effects on autotetraploid genomic prediction. *Crop Science*, 60(2), 656-665.
- Bassi, F. M., Bentley, A. R., Charmet, G., Ortiz, R., & Crossa, J. (2016). Breeding schemes for the implementation of genomic selection in wheat (*Triticum* spp.). *Plant Science*, 242, 23-36.
- Batista, L. G., Mello, V. H., Souza, A. P., & Margarido, G. R. (2022). Genomic prediction with allele dosage information in highly polyploid species. *Theoretical and Applied Genetics*, 135(2), 723-739.
- Bourke, P. M., Voorrips, R. E., Visser, R. G., & Maliepaard, C. (2015). The double-reduction landscape in tetraploid potato as revealed by a high-density linkage map. *Genetics*, 201(3), 853-863.
- Bourke, P. M., Van Geest, G., Voorrips, R. E., Jansen, J., Kranenburg, T., Shahin, A., ... & Maliepaard, C. (2018). polymapR—linkage analysis and genetic map construction from F1 populations of outcrossing polyploids. *Bioinformatics*, 34(20), 3496-3502.
- Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., De Los Campos, G., ... & Varshney, R. K. (2017). Genomic selection in plant breeding: methods, models, and perspectives. *Trends in plant science*, 22(11), 961-975.
- de Bem Oliveira, I., Resende Jr, M. F., Ferrão, L. F. V., Amadeu, R. R., Endelman, J. B., Kirst, M., ... & Munoz, P. R. (2019). Genomic prediction of autotetraploids; influence of relationship matrices, allele dosage, and continuous genotyping calls in phenotype prediction. *G3: Genes, Genomes, Genetics*, 9(4), 1189-1198.
- Endelman, J. B. (2011). Ridge regression and other kernels for genomic selection with R package rrBLUP. *The plant genome*, 4(3).
- Endelman, J. B., Carley, C. A. S., Bethke, P. C., Coombs, J. J., Clough, M. E., da Silva, W. L., ... & Yencho, G. C. (2018). Genetic variance partitioning and genome-wide prediction with allele dosage information in autotetraploid potato. *Genetics*, 209(1), 77-87.
- Endelman, J. B. (2023). Fully efficient, two-stage analysis of multi-environment trials with directional dominance and multi-trait genomic selection. *Theoretical and Applied Genetics*, 136(4), 65.
- Falconer, D. S. (1996). *Introduction to quantitative genetics*. Pearson Education India.

Ferrão, L. F. V., R. R. Amadeu, J. Benevenuto, I. de Bem Oliveira, and P. R. Munoz. 2021. Genomic selection in an outcrossing autotetraploid fruit crop: lessons from blueberry breeding. *Frontiers in Plant Science* 12:676326.

Gaynor, R. C., G. Gorjanc, A. R. Bentley, E. S. Ober, P. Howell, R. Jackson, I. J. Mackay, and J. M. Hickey. 2017. A two-part strategy for using genomic selection to develop inbred lines. *Crop Science* 57:2372–2386.

Gaynor, R. C., G. Gorjanc, and J. M. Hickey. 2021. AlphaSimR: an R package for breeding program simulations. *G3: Genes|Genomes|Genetics* 11:jkaa017.

Gerard, D., L. F. V. Ferrão, A. A. F. Garcia, and M. Stephens. 2018. Genotyping polyploids from messy sequencing data. *Genetics* 210:789–807.

Gorjanc, G., R. C. Gaynor, and J. M. Hickey. 2018. Optimal cross selection for long-term genetic gain in two-part programs with rapid recurrent genomic selection. *Theoretical and Applied Genetics* 131:1953–1966.

Hayes, B. J., P. J. Bowman, A. J. Chamberlain, and M. E. Goddard. 2009. Invited review: Genomic selection in dairy cattle: progress and challenges. *Journal of Dairy Science* 92:433–443.

Jia, C., F. Zhao, X. Wang, J. Han, H. Zhao, G. Liu, and Z. Wang. 2018. Genomic prediction for 25 agronomic and quality traits in alfalfa (*Medicago sativa*). *Frontiers in Plant Science* 9:1220.

Jiang, Y., and J. C. Reif. 2015. Modeling epistasis in genomic selection. *Genetics* 201:759–768.

Kihara, H. 1951. Triploid watermelons. *Proceedings of the American Society for Horticultural Science* 58:217–230.

Meuwissen, T. H. E., B. J. Hayes, and M. E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157:1819–1829.

Muñoz, P. R., M. F. R. Resende Jr., S. A. Gezan, M. D. V. Resende, G. de los Campos, M. Kirst, D. Huber, and G. F. Peter. 2014. Unraveling additive from nonadditive effects using genomic relationship matrices. *Genetics* 198:1759–1768.

Pang, W., Q. Wang, C. Li, W. He, J. Wang, S. Tian, and L. Yuan. 2025. CLPS1 gene-mediated manipulation of 2n pollen formation enables the creation of triploid seedless watermelon. *Molecular Horticulture* 5:48.

R Core Team, . (2021). R: A language and environment for statistical computing. *R foundation for statistical computing, Vienna, Austria*.

Rhodes, B., & Zhang, X. (2008). Watermelon breeding. In J. Prohens & F. Nuez (Eds.), *Handbook of Plant Breeding: Volume 1: Vegetables I* (pp. 1–20). Springer.

Serang, O., Mollinari, M., & Garcia, A. A. F. (2012). Efficient exact maximum a posteriori computation for Bayesian SNP genotyping in polyploids. *PloS one*, 7(2), e30906.

Sverrisdóttir, E., Sundmark, E. H. R., Johnsen, H. Ø., Kirk, H. G., Asp, T., Janss, L., ... & Nielsen, K. L. (2018). The value of expanding the training population to improve genomic selection models in tetraploid potato. *Frontiers in plant science*, 9, 1118.

Sybenga, J. (2012). *Meiotic configurations: a source of information for estimating genetic parameters*. Springer Science & Business Media.

Uitdewilligen, J. G., Wolters, A. M. A., D'hoop, B. B., Borm, T. J., Visser, R. G., & Van Eck, H. J. (2013). A next-generation sequencing method for genotyping-by-sequencing of highly heterozygous autotetraploid potato. *PloS one*, 8(5), e62355.

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *Journal of dairy science*, 91(11), 4414-4423.

Vitezica, Z. G., Varona, L., & Legarra, A. (2013). On the additive and dominant variance and covariance of individuals within the genomic selection scope. *Genetics*, 195(4), 1223-1230.

Voorrips, R. E., Gort, G., & Vosman, B. (2011). Genotype calling in tetraploid species from bi-allelic marker data using mixture models. *BMC bioinformatics*, 12(1), 172.

Wehner, T.C. (2008). Watermelon. In: Prohens, J., Nuez, F. (eds) *Vegetables I. Handbook of Plant Breeding*, vol 1. Springer, New York, NY. https://doi.org/10.1007/978-0-387-30443-4_12

Wickham, H. (2016). Data analysis. In *ggplot2: elegant graphics for data analysis* (pp. 189-201). Cham: Springer international publishing.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics bulletin*, 1(6), 80-83.