GWAS with GBLUP on *C. elegans* for Locomotion Behavior

Alix Boutheroue-Desmarais

14 February to 28 June

Tutors

Henrique Teotonio
François Mallard

INSTITUT DE BIOLOGIE DE L'ECOLE NORMALE SUPERIEURE (IBENS)

Master de Bio-informatique, Sorbonne Université

Contents

1 Abstract

Multitrait approaches in genome-wide association studies (GWAS) are essential for understanding the genetic architecture of complex phenotypes, as they account for the genetic and environmental correlations between traits. These methods enhance the detection of quantitative trait loci (QTL) involved in phenotypic variation by integrating information across multiple traits. In our study, we employed the genomic best linear unbiased prediction (GBLUP) method within a GWAS framework to investigate the genetic basis of locomotion traits in the *Caenorhabditis elegans* Multiparental Experimental Evolution (CeMEE) Panel. While the univariate application of GWAS by GBLUP has been shown to be both computationally efficient and accurate, the consistency and effectiveness of its multivariate application had not yet been demonstrated. Our results reveal that pruning does not significantly impact the analysis, the construction of genomic relationship matrices (GRM) yields consistent results, and multivariate GWAS using GBLUP identifies SNP associations that are coherent with those found in univariate approaches, in addition to uncovering more associations. These findings tend to show the consistency of multitrait models and their efficacy in capturing the complexity of genetic influences on phenotypes.

2 Introduction

Organisms cannot be considered as the sum of independent phenotypes, as they are the result of traits that are linked both genetically and environmentally. Genetic correlation between traits can have several origins, including pleiotropy, where one locus or gene is associated with several traits, or where traits share biological pathways or other more indirect links. Consequently, multivariate phenotype studies that take account of this covariance between traits are essential for understanding complex phenotypes. Indeed, the highly polygenic nature of these phenotypes poses an immense challenge to understanding the biological mechanisms that link individual genetic variants to phenotypes.

The Genome Wide Association Study (GWAS) aims to address this challenge. Using a genome-wide approach, GWAS examines thousands of genetic variants in a population to identify those associated with specific traits. This methodology makes it possible to detect quantitative trait loci (QTL) involved in phenotypic variation. By analyzing the correlation between genetic variations and observed traits, GWAS helps to uncover the genetic basis of complex phenotypes and to understand how specific combinations of variants influence these traits [1]. GWAS approaches have been extensively used since the beginning of the 20th century and several methods have been developed. Here, in order for the reader to understand the choices made to test for association, it seems important to begin by exposing the theoretical framework of GWAS [2]: Fisher postulates that continuous traits have both environmental and genetic components $y_i = \sum_{k=1}^{K} Z_{ik} + \varepsilon_i$, with ε_i being the environmental effect, Z_{ik} the effect of Mendelian factor k on subject i. Nowadays, we can gather observations on Z_{ik} and introduce a matrix $X = x_{ij}$ with $x_{ij} \in 0, 1, 2$ depicting the genotypes at bi-allelic markers and introduce it into Fisher's model: $y_i = \sum_{k=1}^{K} \beta_k X_{ki} + \varepsilon_i$ with β corresponding to the SNP effect. Most of the models used today are using a SNP by SNP method, mixed linear models are defined for each SNP as:

$$Y \sim W\alpha + X\beta + g + e \tag{1}$$

$$g \sim \mathcal{N}(0, \sigma_A^2 A) \tag{2}$$

$$e \sim \mathcal{N}(0, \sigma_e^2 I) \tag{3}$$

where, for each individual, Y is a vector of phenotype values, W is a matrix of covariates including an intercept term, α is a corresponding vector of effect sizes, X_s is a vector of genotype values for all individuals at SNP s, β_s is the corresponding fixed effect size of genetic variant s (also known as the SNP effect size), g is a random effect to estimate SNP-based heritability, A measures the additive genetic variation of the phenotype, A is the standard genetic relationship matrix, σ_e^2 measures residual variance, and I is an identity matrix.

However, this kind of modeling doesn't match (2), leading to several consequences:

- It fails to accurately determine the degree of relatedness among individuals. Using (1), the effects of unmodeled genetic factors are included in the error term, causing correlations that can mislead the significance evaluation of each β_k value.

- Omitting relevant variables biases the estimates of the β_k values, potentially missing important loci. Consequently, SNPs identified in single SNP regressions explain only a small portion of trait variance, contributing to the missing heritability problem.

Moreover, there is an important statistical problem: the number of SNP effects is usually much larger than the number of observed phenotypes. One solution to these problems is to model SNP effects as random effects and make prior assumptions about the variance explained by their effects. The genomic best linear unbiased prediction (GBLUP) method assigns the same variance to all loci, fits all SNP effects simultaneously [3], and is able to consider multiple traits 4. GBLUP allows us to obtain breeding values, which are estimates of an animal's genetic merit, by using genomic information summarized in a genomic relationship matrix (GRM) for traits for each line. We then use back-solving methods to deduce the SNP effects [3] [4]. However, this backsolving step has not yet been demonstrated for multivariate analyses, and the validity of this analysis remains to be explored.

Another advantage of GWAS by GBLUP [5] is its Bayesian approach over the classic frequentist GWAS. The frequentist approach is a statistical school of thought in which inferences about unknowns are justified not with reference to probabilities for the inferred value, but on the basis of measures of performance under imaginary repetitions of the procedure that was used to make the inference. In contrast, Bayesian statistics is a school of thought that holds that inferences about any unknown parameter or hypothesis should be encapsulated in a probability distribution, given the observed data. Computing this posterior probability distribution usually proceeds by specifying a prior distribution that summarizes knowledge about the unknown before the observed data are considered, and then using Bayes' theorem to transform the prior distribution into a posterior distribution [6]. This prior enables the incorporation of information about genetic effect sizes. This can enhance estimation accuracy, especially in cases where prior biological knowledge or previous studies suggest certain genetic architectures [7]. Moreover, the Bayesian framework provides access to posterior distributions [4], offering a probabilistic interpretation of the results. It also allows for the computation of the G-matrix, the additive genetic variance-covariance matrix of multiple traits, with the variance between traits on the diagonal and the covariance otherwise. Previously, this matrix had been computed only with phenotypes [8] [9].

Another reason to use multivariate models is that, in many cases, they provide more power to GWAS analysis. The settings in which multivariate analyses have increased power can be counter-intuitive. In particular, they have power advantages even when only one of the phenotypes is associated with genotype, a setting that might naively be expected to favor univariate analysis [10].

Locomotion is a complex phenotype that can be decomposed into six independent traits [8], and previous works shows it is under selection [9]. Then identifying loci allows us to map loci under selection, which is valuable for studying evolutionary forces. We used the *Caenorhabditis elegans* Multiparental Experimental Evolution (CeMEE) Panel [11] [12] to conduct our GWAS. We want to explore the polygenic structure of locomotion behavior measured in two environments and ask whether traits correlations within and between environments are due to genetic linkage between QTLs loci, pleiotropic QTLs or a combination of both. The first step was to call variants for our sequenced lines. Results were compared between each population we considered, and heterozygosity was explored. In the second step, we studied the impact of genetic relationship matrix (GRM) construction and pruning on the GBLUP model. Our main goals were to compare different methods for constructing the kinship matrix (GRM), evaluate the consistency and benefits of multivariate GWAS using the GBLUP approach compared to univariate methods, and investigate pleiotropy and linkage in multiple traits and across different environmental conditions.

The results obtained represent a first step towards demonstrating the feasibility of using GWAS by GBLUP in a multivariate context. Additionally, they allow the identification of QTLs associated with locomotion, a trait under selection.

3 Material and Methods

3.1 Experimental Evolution (EE)

The fundamental premise of Experimental Evolution (EE) is to cultivate populations within controlled environments to investigate the evolution of their genotypes and phenotypes. These populations are typically compared either to an ancestral group to observe divergence or to a control group for differentiation analysis. Through EE, researchers gain insight into both the ancestral state and the evolutionary processes that have shaped these populations. In the context of EE, *C. elegans* is a relatively new but promising model organism. Its short life cycle of about four days makes long-term evolutionary studies feasible. Additionally, *C. elegans* can be easily and reliably cryopreserved, ensuring the stability and reproducibility of experiments [13].

3.2 *C. elegans*

Since the nineties, the interest in *Caenorhabditis elegans* (*C. elegans*) has been recognized not only for its ease of use in various biological fields. To understand this report, some basics of *C. elegans* biology are needed. They have a rare androdioecious reproduction system. Hermaphrodites in this species are capable of selfing and outcrossing with males but not with other hermaphrodites. This unique system allows researchers to manipulate sex determination and achieve populations with variable ratios of males, females, and hermaphrodites, thus facilitating different degrees of selfing and outcrossing [13].

C. elegans exists in two sexes, distinguished genetically by their X-chromosome complement: XX worms are hermaphrodites, while XO worms are males. There is no sex-specific chromosome like a Y chromosome. They have five autosomes. Their chromosomes are holocentric, possessing multiple kinetochores along their length rather than the single centromere typical of other chromosomes.

3.3 The CeMEE Panel

The host laboratory is therefore doing EE on *C. elegans*, focusing on 16 natural lines of *C. elegans* (AB1, CB4507, CB4858, CB4855, CB4852, CB4856, MY1, MY16, JU319, JU345, JU400, N2 (ancestral), PB306, PX174, PX179, and RC301). From these 16 founder populations, they evolved several populations. Firstly, A6140: the 16 founders were mixed, then their progeny were replicated 6 times and these 6 replicates evolved for 140 generations under androdioecious conditions. Based on this population, numerous experiments and new populations were evolved, including : CA[1-3]50 and CA[1-3]100: Control androdioecious populations evolved over 50 and 100 generations, respectively. GA[1,2,4], GT[1,2], and GM[1,3]: Androdioecious, trioecious, and monoecious populations evolved in gradually increasing NaCl concentrations. EEV: Derived from

A6x140 with introgression of GFP, affecting the recombination rate. LR: Introgression in the rec-1 gene, impacting the distribution of meiotic crossovers. SMR: Control for introgression evolving in NaCl concentration. These populations were evolved in a constant environment after a domestication period of 140 generations. The populations were cultured at a constant size of 10,000 with an expected effective population size of 1,000. Recombinant Imbred Lines (RILs) have been derived from these populations, generated by selfing hermaphrodites for at least 10 generations. The CeMee panel is made up of over a thousand of sequenced RILs. [11] [12]

3.3.1 Data

For our GWAS, we considered a set of 1,348 sequenced lines, which had been sequenced during previous experiments at the lab [11][12][14], and some were sequenced newly at a New York laboratory. All lines were already mapped to the reference genome WS283. During CeMee experiments [11] [12], 273 A6 RILs, 298 CA, 162 GA, 100 GM an 88 GT were sequenced, from experiment on recombination rate [14] 92 LR, 100 SMR and at the NYC laboratory 47 lines from A6, 186 CA. Lines were sequenced at different resolutions, resulting in varying sequencing and mapping quality (Supplementary Figure S1). Lines sequenced at the NYC laboratory exhibit significantly higher quality compared to others. Some lines experienced quality issues, losing more than 50% of their initial reads after the mapping steps. This suggests potential contamination, such as bacterial contamination, which is the food source for the worms.

Variant calling is the process by which we identify variants from sequence data. One of the aims of the laboratory is to add the CeMEE panel to the Caenorhabditis Natural Diversity Resource (CaeNDR) [15]. Our lines are compared to the reference genome WS283, and GATK v4.5.0.0 is used to call variants. Preprocessing steps were done with MarkDuplicates, AddOrReplaceGroups, BaseRecalibrator, and ApplyBSQR. Variant calling steps used HaplotypeCaller, CombineGVCFs, and GenotypeGCVFs. The filtering step used both VariantFiltration and bcftools 1.10.2 to perform hard-filtering (i.e., remove variants that are filtered) with the thresholds set in the CaeNDR pipeline. Sites were flagged as PASS if: Variant quality (QUAL) > 30 || Variant quality normalized by read depth (QD) > 20 || Strand bias of ALT calls: strand odds ratio (SOR) < 5 || Strand bias of ALT calls: Fisherstrand (FS) < 100 || Fraction of samples with missing genotype < 95% || Read depth (DP) > 5 || Site is not heterozygous. In order to refine our filtration step, we created simulated data in which the positions of real variants are known and examined each metric independently and in combination. Moreover, the HaplotypeCaller function requires a file with known hyper-variable regions that are not included in the variant calling. Our goal was to obtain as many markers as possible, and a blank file was used. This practice is discouraged by GATK, and another approach to address this issue should be considered. At this step, 655,806 SNPs were found for the 16 founder lines. SNPs of the founders can be considered as the standing genetic variation of our populations, because all SNPs that appear afterward can be considered as mutations. SNPs of all lines could be due to standing genetic variation or new mutations, so they need to be analyzed and compared to the founders' SNPs in order to determine their origin. At the end of the GATK pipeline, 551,241 SNPs were called for CeMee panel's lines, founders exclude. The proportion of SNPs per chromosome vary between 0.2% and 0.8% (describes in Figure S2).

3.3.2 Isotypes

However, as mentioned before, some lines in the CeMEE panel were sequenced twice and need to be removed. Moreover, as explained in a previous lab paper [12] and in CaeNDR, some lines can be very similar to each other, referred to as isotypes, and only one exemplar needs to be kept. Finding isotypes was not an easy task. The first step was to compute the concordance matrix, which compares each pair of lines. The concordance score is calculated as $\frac{\text{Number of shared SNPs}}{\text{Number of known SNPs}}$. Some lines with quality issues, leading to many NAs, were removed at this step (68 lines). Afterward, a cutoff was chosen by looking at the concordance scores of lines known to be present twice; the cutoff was set at 0.9945. Known isotypes with concordance scores lower than this cutoff were removed (48 lines). The next and final step was to find isotype groups. Each line with a concordance score greater than the cutoff was considered a *"friend"*. We compiled a list of these *"friends"* and sought to find true *"friends"* among them, meaning groups of lines where they are all *"friends"* with each other. This reciprocal part was the hardest to compile. To do this, we considered a matrix for each group of *"friends"*, with each *"friend"* as a row and column, 0 if the concordance score is lower than the cutoff and 1 if it is greater. Each square around the diagonal represents a group of true *"friends"*, and to seek isolated groups of *"friends"*, each of these groups of true *"friends"* was compared with other groups.

A total of 390 lines were involved in 135 isotype groups, and 255 lines were removed at the end of this step, leaving 959 lines. The representative line is defined as the one with the fewest missing values and the best quality. There are groups of isotypes between lines that are not from the same population, which may indicate contamination, data entry errors, or experimental errors. Further analysis is required to determine where in the experiment these errors occurred.

3.3.3 Heterozygosity

Filtering lines To ensure the quality of our analysis, we removed lines with too much heterozygosity; the cutoff was set at 0.085, resulting in the removal of 15 lines. This left us with 944 lines. The CeMEE panel thus contains 944 lines, distributed as follows: 233 lines from A6, 310 from CA, 134 from GA, 9 from GM, 79 from GT, 85 from LR, and 92 from SMR. To conduct the GWAS (Genome-Wide Association Study), we selected only phenotyped lines, reducing the total to 747 lines: 178 A6, 261 CA, 134 GA, 82 LR, and 91 SMR. The phenotypes measured for the isotype lines removed were reassigned to the line representing the group. Alternative methods, such as those used in the CaeNDR database, could conserve information at these loci by transforming them into probable homozygotes. Another approach would have been to remove problematic SNPs if they were few in number.

Characterizing heterozygosity In the variant calling pipeline, heterozygous loci are marked with the 'is_het' flag, allowing them to be turned into missing values later. Tagging heterozygous loci allows heterozygosity to be traced along chromosomes. We define windows of 500 SNPs, where heterozygosity is calculated as the average number of heterozygotes in each window for each individual. To better understand the origin of these regions, we used the hyperdivergent regions (HDRs) of the founders known in the CaeNDR database [15], HDR are defined as segments of the genome with a much higher than average rate of variation. If an HDR region overlaps our SNP window, it is marked as hyperdivergent (open interval).

To test the association between HDR regions and heterozygous regions, we calculated the proportion of the 5% of windows with the highest heterozygosity associated with HDRs. The null distribution was obtained by mixing the heterozygosity-HDR association and calculating the proportion obtained, and repeating this operation 50,000 times to plot the distribution we would have obtained by chance. To avoid bias from peaks of heterozygosity, we performed this randomization per chromosome for the A6 population.

3.3.4 Imputation

Imputation is the statistical inference of unobserved genotypes. It is commonly performed in genome-wide association studies because it greatly increases the number of markers that can be tested for association with a trait. Beagle 5.0 imputation is based on identity by state (IBS). Genotyped chromosome segments are compared between the haplotype to be imputed and the reference haplotype in order to identify long IBS segments. When an IBS segment is found, ungenotyped alleles are inferred from the reference haplotype. To produce a posterior distribution for each possible allele at an imputed marker, a hidden Markov model is used.

Imputation of the founders gave us a set of 524,697 SNPs before filtration, and 501,776 after, results for all populations can be seen in Figure 1. Imputation can infer heterozygous loci and multi-allelic ones, which must be removed during the filtration step. Filtration switches heterozygous sites to missing ones and removes mono-morphic, multi-allelic, and SNPs with a minor allele frequency less than 0.05 (MAF filtering doesn't remove SNPs). The imputation was done by population for all the lines. Even though populations should be quite equivalent, this approach allows for better qualification of the particularities of each population. However, filtering of variant calling was done at the population level. The imputation stage proved extremely effective. Figure 1 shows the logarithm of the average missing values per line in each population before and after imputation. The gain is considerable, with virtually no missing values in each population after imputation. GM

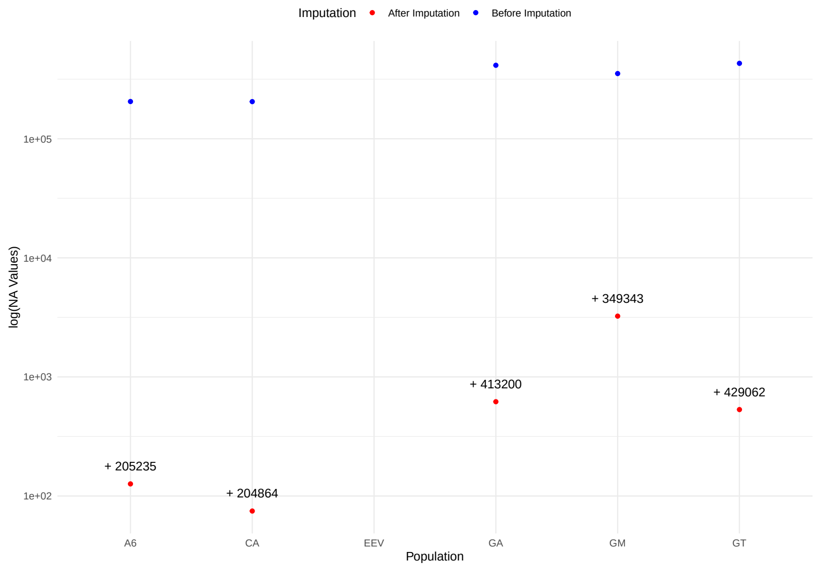

Figure 1: Number of SNPs imputed during imputation

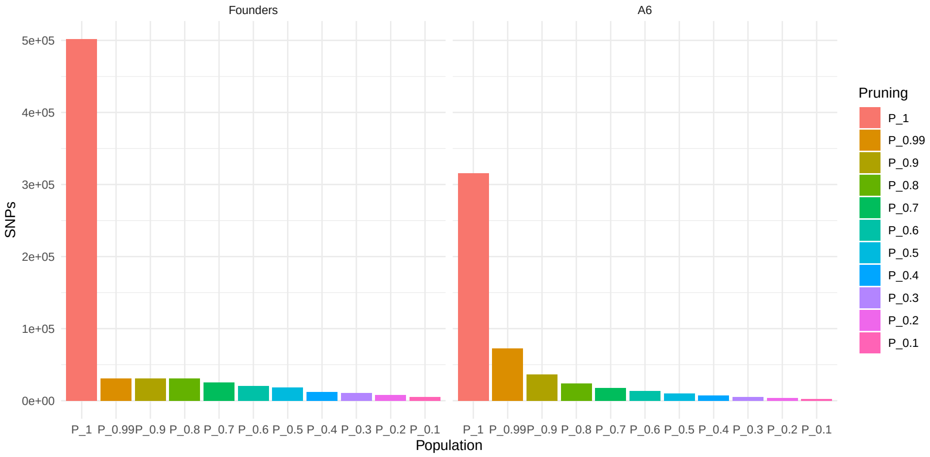

Figure 2: Pruning Results

has the most missing values after imputation, which can be easily explained by the low number of lines kept in this population. Indeed, no reference haplotype was given here, so the algorithm uses haplotypes in the population to infer unknown positions. Thus, the number of lines and similarity between them are proportional to the results of imputation.

3.3.5 Pruning

SNP pruning aims to subset them and keep only independent SNPs, i.e., those in approximate linkage equilibrium (LD). As with the choice of the kinship matrix, the method of pruning impacts results. To prune SNPs, we used plink v1.9. The method implemented in this package uses the strength of LD between SNPs within a specific window (region) of the chromosome and selects only SNPs that are approximately uncorrelated, based on a user-specified threshold of LD. It uses the first SNP in genome order and computes the correlation with the following ones (e.g., 50). When it finds a large correlation, it removes one SNP from the correlated pair, keeping the one with the largest minor allele frequency (MAF), thus possibly removing the first SNP. It then continues with the next SNP not yet removed. Ten different R-squared (i.e., correlation coefficient between pairs of SNPs) cutoffs of LD (0.99, 0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3, 0.2, 0.1) were set to prune markers, and the pruning was conducted in a 200-kb sliding window with 10 variants. The results are shown in Figure 2 for A6 and founders, they are representative of other population (All results in S3). As expected, the number of SNPs decreases significantly between the non-pruning dataset (P_1) and pruning with a threshold of 0.99. There is a plateau at the first pruning thresholds for founder because they are containing a low number of lines (same results for GM). Their loci show either very strong or weak linkage, and the limited number of lineages does not allow these distinctions to be explored in detail.

3.4 Locomotion

Inbred lines from the experimental populations were phenotyped at two different lab locations by several experimenters. The locomotion behavior was measured using the Multi-Worm Tracker (MWT). MWT detects and loses objects over time as individual worms enter and leave the field of view or collide with each other. Males and hermaphrodites do not move the same way and are therefore considered in different models. Here, we consider only hermaphrodites. In a one-dimensional space, individual locomotion behavior can be described by the transition rates of activity and direction. The expected sex-specific transition rates between forward, still, and backward movement states are modeled with a continuous-time Markov process. Only six of the nine possible transitions are independent, and only these will be considered, such as the transitions from forward to still (FS), backward to forward (BF), and so on. The phenotypic data were obtained and the transition rates calculated previously in the lab [8].

Two conditions were considered: the classic condition, named NGM, and the salt condition, which is a stressful condition. The six traits were measured with at least two replicates per individual and were centered. Differences between conditions varied according to the populations studied.

Population A6 : Traits FS, BS, and SB were similar between conditions, while SF, BF, and FB had higher values under salinity conditions. **CA population:** All traits vary except SB and FB. Traits SF and BF show similar trends to A6. FS and BS show higher trends in NGM. **Population GA :** Traits SF and BF show similar behaviour to A6. FS and BS have higher trends in NGM, SB and FB have lower trends in NGM. **LR population:** FS, BS, and SB are higher in NGM. SF, BF, and FB are lower in NGM. **SMR population:** A bimodal distribution was observed for SF and BF, with SF predominating during the specific period and BF less marked, this underlight the fact that we need to take covariables, such as the period of the experiment, in account to correct

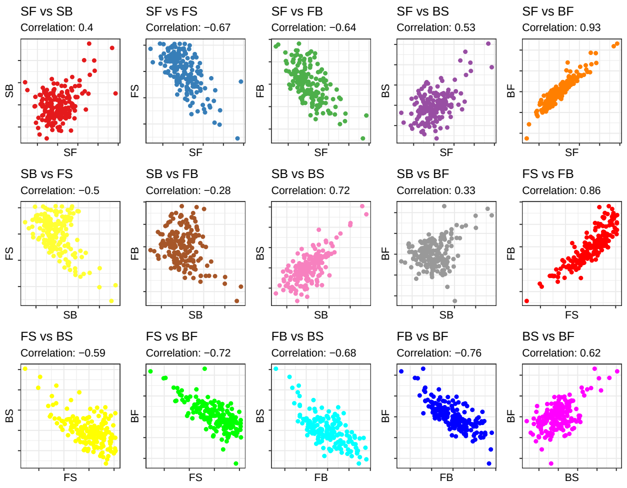

Figure 3: Correlation between traits for phenotypic data

the phenoypic measures. FS, BS, SB, and FB are higher in NGM, while SF and BF are lower in NGM. No population has a trait evolution similar to another. We note that the BF and SF traits systematically evolve in the same direction, as do FS and BS. This can be seen as an expression of correlation between traits, to explore it, Pearson correlation coefficients are used to measure the strength and direction of the linear relationship between two quantitative variables (Figure 3). For all pairs of traits except SF - SB, SB - FB, and SB - BF, we observed absolute correlations greater than or equal to 0.5, so there is a strong correlation between these six traits.

3.5 GWAS by GBLUP

GWAS by GBLUP is a hybrid approach that enhances traditional GWAS by incorporating the genomic prediction capabilities of GBLUP, thereby improving the identification and understanding of the genetic architecture of complex traits. Unlike the classical GWAS approach, where marker effects are viewed as fixed, in GWAS by GBLUP, marker effects are first viewed as random. They are assuming to come from a normal distribution. Therefore, some prior has to be set to estimate these effects. The introduction of priors has the effect of "regressing" the estimators towards the a priori values, a process known as shrinkage. This kind of regularization allows us to address the $n >> p$ problem, where the number of markers (p) greatly exceeds the number of observations (n) [4].

3.5.1 Genomic Relationship Matrices

Genomic Relationship Matrices (GRMs) offer a valuable approach in genomic evaluations due to their ability to capture genetic relationships among individuals using shared DNA markers. Unlike traditional pedigree-based methods, GRMs leverage information from genetic markers to estimate additive genetic relationships more accurately. The key advantage of GRMs lies in their capacity to exploit the variation in relationships among individuals. Instead of assuming fixed relationships as in pedigree-based approaches, GRMs allow for a better control for allele frequencies and linkage disequilibrium across the genome [3].

Population structure and kinship are widespread confounding factors in GWAS, and GRM is a method for simultaneously accounting for population structure and kinship. Because of experimental evolution (EE), we are aware of the population structure of our samples, which is a particular concern in human genetics where individuals evolve over long-term periods. In our study, we only consider short-term evolution, making GRM a good approximation for our analysis. In this Bayesian approach, the genetic relationship matrix is used to "shrink" the individual SNP effects, reducing the impact of noise and enhancing the stability of the predictions.

The genomic best linear unbiased prediction (GBLUP) model for a trait builds a matrix of relationships between all individuals based on genomic data and uses it to partition the phenotypic variance into components, thus predicting breeding values. Assumptions made in methods for constructing the relationship matrix directly affect the accuracy of the results. There are various methods for constructing relationship matrices using genomics.

To perform a GBLUP analysis, it is essential to use a kinship matrix that is invertible and positive definite. A matrix is said to be invertible if it has a unique inverse such that the product of the matrix and its inverse gives the identity matrix. A matrix is positive definite if all its eigenvalues are strictly positive, guaranteeing that all variances calculated from this matrix are positive. These conditions are guaranteed by the R packages *MASS* and *Matrix*. The condition number of the matrix is a measure of the numerical stability of the solutions of a system of linear equations associated with this matrix. A high condition number indicates that the matrix is ill-conditioned, meaning that small perturbations in the data can lead to large variations in the solutions, making the calculations unstable and unreliable, and potentially the matrix is stably non-invertible. In this case, the pseudo-inverse is calculated using the `ginv` function from the *MASS* library. If the matrix is well-conditioned (low condition number), the classical inverse is calculated using the `solve` function. Once the inverse or pseudo-inverse has been obtained, we check whether this matrix is positive definite by examining its eigenvalues using the `eigen` function in the *Matrix* library. If the inverted matrix is not positive definite, the diagonal values are adjusted by gradually adding a small epsilon until all the eigenvalues are strictly positive. This process ensures that the matrix used for the GBLUP analysis meets the necessary criteria of invertibility and definite positivity.

Instead of using the traditional relationship matrix from the pedigree in the *ginverse* option of the MCMCglmm package, the inverse of the GRM matrix was used.

VanRaden The estimate of GRM by VanRaden [16] has been defined as follows: first, M is defined as an incidence matrix that specifies which alleles each individual has inherited, coded as -1, 0, 1 or 0, 1, 2. Let the frequency of the second allele at locus i be p_i, and then consider the matrix P such that its column i is $2(p_i - 0.5)$. Z is the subtraction of P from M, which sets the mean values of the allele effects to 0. Let's denote the GRM as:

$$A_{VR} = \frac{ZZ'}{2\sum_i(1-p_i)}$$

The division by $2\sum_i(1-p_i)$ scales A_{VR} to be analogous to the numerator relationship matrix from pedigree data.

The GRM from VanRaden was accessed using the *estimGenRel* function from the R package rutilstimflutre. The VanRaden approach assumes all loci to be at the Hardy-Weinberg Equilibrium (HWE), meaning that from generation to generation, allele frequencies in a population remain constant. These assumptions influence the orthogonality of the estimates of the components of the genetic variance (e.g., additive, dominance, etc.) and, hence, whether the estimates change when other components are included in the model. Moreover, VanRaden doesn't account for the linkage disequilibrium (LD).

Natural and Orthogonal Interactions Approach (NOIA) Genotypic values consist of additive, dominance, and epistatic components. Classical genetic theory assumes ideal conditions like random mating and Hardy-Weinberg equilibrium (HWE), which are often not met in practice. The NOIA approach [17] addresses this by using genotypic frequencies to construct incidence matrices, relaxing the HWE assumption.

Cockerham's (1954) epistatic model breaks gene variance into orthogonal components. For two loci, A and B, eight orthogonal contrasts $w_{1,\dots,8}$ represent additive, dominance, and epistatic effects. These contrasts allow the partitioning of genetic variance into independent components, facilitating the modeling of specific genetic effects without interference from other effects.

Genotypic values for two loci j and k are expressed as:

$$g_{jk} = [1, w_1, w_2, w_3, w_4, w_5, w_6, w_7, w_8]\theta$$

with $\theta' = [\mu; a_j; d_j; a_k; d_k; (aa); (ad); (da); (dd)]'$. Orthogonality ensures no genetic covariance between components. The kinship matrix separates genotypic values into additive (g_A) and dominance (g_D) components. Thus, $g = \mu + g_A + g_D$ with $g_A = h_a a$ and $g_D = h_d d$. When the linkage equilibrium (LE) assumption is satisfied, the orthogonal properties described above are found.

The GRM from NOIA was computed using the *estimGenRel* function from the R package rutilstimflutre.

Hoffman To simplify the notation, this method of estimating the GRM will be referred to as Hoffman [18]. Similar to the previous approach, population structure is treated as a random effect. Considering X as the genotype matrix ($n \times p$), the GRM is defined as $A_h = XX^T$. However, if we consider the singular value decomposition (SVD) underlying principal component analysis (PCA), which treats population structure as fixed, we obtain $X = USV^T$, where the first i principal components are the first i columns of U $(n \times n)$. S $(n \times n)$ is a diagonal matrix such that $S \approx \text{diag}(s)$ with s containing the singular values corresponding to each principal component, and V $(p \times n)$ contains the loadings on each marker. Each marker in X has been centered on the mean and scaled. The GRM can then be written as $A_h = USV^T(USV^T)^T = USV^TVS^TU^T = US^2U^T$, where $S^2 = \text{diag}(s^2)$. This formulation demonstrates the relationship with PCA, as the principal components captured by U and the singular values in S provide a means to account for population structure within the GRM. Our own function is used to compute this matrix.

3.5.2 GBLUP

GBLUP was developed from BLUP to incorporate genomic information, a similar method is SNPBLUP (or RRBLUP), which uses the model

$$y = 1_n\mu + \sum_i Wq_i + e,$$

where μ is the mean, W is a matrix of genotypes coded as 0, 1, or 2, and q_i is the effect of each SNP. While SNPBLUP estimates SNP effects directly, GBLUP reduces the dimensions of genetic effects in mixed model equations from $m \times m$ (number of markers) to $n \times n$ (number of individuals), making it more computationally efficient. Theory shows GBLUP and RRBLUP are equivalent if the number of QTL is large, no major QTL exist, and QTL are evenly distributed [3].

GBLUP assumes equal variance for all marker effects, suitable for traits following the infinitesimal model. For traits controlled by major genes, Bayesian methods with shrinkage priors or GWAS can identify causal variants effectively, which can then inform the construction of a weighted genomic relationship matrix (G). GBLUP is reliable in both single-trait and multi-trait approaches [4].

Our six transition rates were fitted as a multivariate response variable Y in the model:

$$y = Xb + Wr + Zg + e,$$

where X is a design matrix for fixed effects, b is a vector of fixed effects, W relates to random effects, r is a vector of random effects, Z allocates records to genetic values, g is a vector of additive genetic effects, and e is a vector of random normal deviates with variance σ_e^2. Additionally, $\text{var}(g) = G\sigma_g^2$, where G is the genomic relationship matrix and σ_g^2 is the genetic variance. Temperature, relative humidity, and density are fixed effects, while block effects are random.

We used the R package MCMCglmm for modeling, with priors as the matrix of phenotypic variances for each trait. Model convergence was verified visually and by ensuring autocorrelation remained below 0.05. We used 10,000 burn-in iterations, a thinning interval of 100, and a total of 110,000 MCMC iterations.

G-matrix The G matrix, or additive genetic variance-covariance matrix, describes the additive genetic variability for several traits and the genetic covariances between these traits. Conceptually, selection on several traits can be imagined as a multidimensional surface where each dimension represents a specific gradient of strength and direction of selection for each trait and the interactions between them. Responses to selection depend on the size and shape of this G matrix [9]. In our study, we used a model similar to that in the previous article on locomotion by F. Mallard [9] [8], with one key difference: he was using a BLUP model, whereas we included a kinship matrix and used a GBLUP model. This allows us to compare the impact of the kinship matrix on the G matrix against a model that considers the identity matrix as the GRM.

3.5.3 Back-solving

Breeding value is a measure of an individual's genetic potential to pass on traits to its offspring, typically estimated using genetic information and statistical models. It represents the sum of the average effects of all genes an individual carries that can be transmitted to the next generation. Let g be the vector of breeding values; they are a linear function of SNP effects $g = Zu$ where Z is a matrix relating genotypes of each locus, and u is a vector of SNP marker effects. Then the variance of SNPs is:

$$\mathrm{Var}(g) = \mathrm{Var}(Zu) = ZZ'\sigma_u^2 = G^*\sigma_g^2.$$

However, ZZ' is equivalent to the GRM, which means G^* in the previous equation, and σ_u and σ_g are equal [19].

The SNP effects can be found with the following linear system:

$$u = M'(MM')^{-1}g,$$

where $(MM')^{-1}$ is the generalized inverse, and M is the incidence matrix for SNP effects based on the SNP genotypes used in the A_{VR} matrix definition, as previously presented A, the GRM [20] [21].

The equation presented here is for univariate analysis, considering only the variance of the SNPs. How to deal with covariance in this model has not been properly demonstrated yet, and the relevance of backsolving for multivariate cases has to be demonstrated.

3.5.4 Credibility Interval

The credibility interval corresponds to the range of values around the central estimate of a parameter, constructed using Bayesian methods. This interval represents the uncertainty associated with the estimate. The wider the interval, the greater the uncertainty. For both breeding value and SNP effects, the latter are of particular interest when their effect is significantly different from zero. We can then define the credibility interval, set at 95%, which means that there is a 0.95 probability that

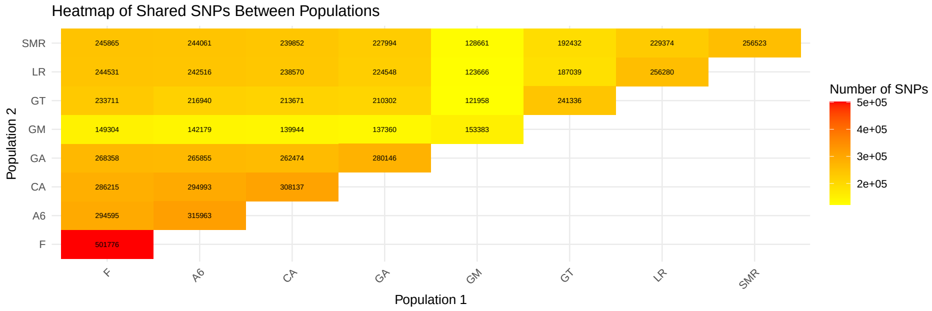

Figure 4: Heatmap of shared SNPs between populations

the effects lie between the extreme values of this interval. If this interval does not overlap with zero, there is a 95% chance that the effects are significantly different from zero and therefore have an impact.

4 Results and Discussion

4.1 SNPs

In order to assess the origin of SNPs found in each population, we plotted Figure 4, which depicts the SNPs shared across populations and founders. We observe that the founders have the largest number of SNPs, and all the populations share a large number of SNPs in common with them. This is expected due to standing genetic variation, which allows for short-term adaptation to the environment. However, the populations also have SNPs absent in the founders, around 20,000 for A6. Although these variations could result from mutations, we consider them due to temporal constraints; however, according to our hypotheses, they should not be taken into account. The A6 population, the most ancestral of the evolved populations, has the highest number of SNPs. The hypothesis is that the number of SNPs decreases over time as a result of genetic drift and selection. Overall, all the populations share the most SNPs with A6, which is expected. GM, with only nine individuals, is the population sharing the fewest SNPs with the others, probably due to the small number of individuals considered, which may lead to under-detection of the SNPs present.

4.2 Heterozygosity

Heterozygosity along chromosomes for the A6 population is plotted in Figure 5 **A**. Black points represent the hyperdivergent regions (HDR), defined in section **3.3.4**. Isolated points can easily be viewed as errors, whereas a region is harder to interpret. Two peaks of heterozygosity are observed on chromosome 5 between positions 16,813,367 and 17,627,977 and align with HDR. This observation is consistent across all the populations studied (Figure S4), with regions of heterozygosity varying in intensity. Several hypotheses can be made: the persistence of heterozygosity could in-

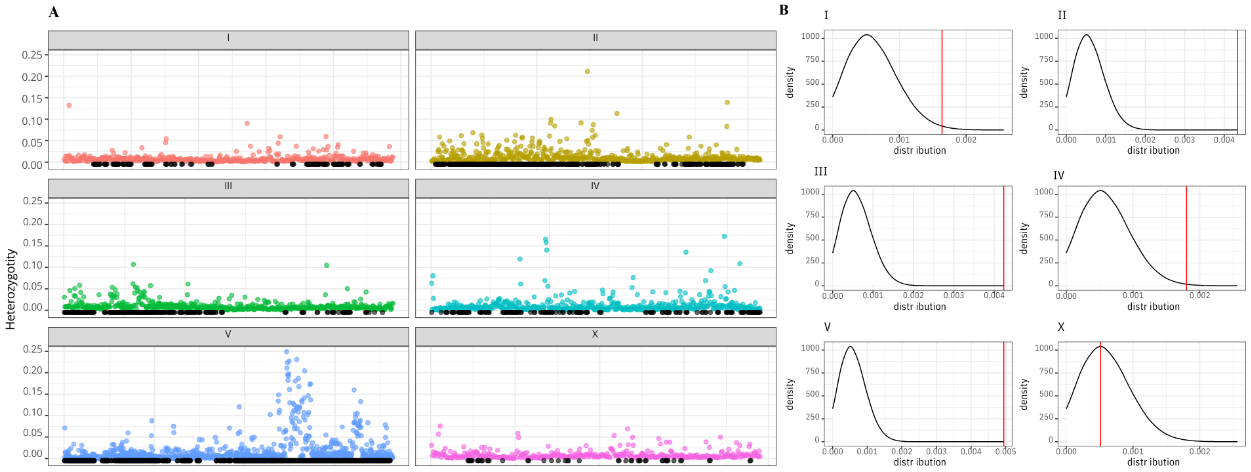

Figure 5: **A**: Heterozygosity along the chromosome for population A6140
B: Comparison of the observed proportion of heterozygous loci associated with HDR (red) to the null distribution.

dicate that heterozygotes have a greater impact on fitness than homozygotes, where both alleles should have equivalent fitness. However, due to self-fertilization, this mechanism shouldn't exist unless the homozygote is lethal. Which is unlikely because it would have a high cost on fitness and then are likely to be eliminate from population. This observation may also be an error of mapping, as the region concerned may have diverged from the reference genome, preventing correct alignment and causing this peak in heterozygosity. The existence of paralogs aligned in the same place could also explain it. In any case, the region considered is very large, making a deletion or duplication of all of it improbable. Further studies are needed to understand this better.

In Figure 5 **C** is depicting the statistical association between HDR and heterozygotes. The observed proportion (red), except for chromosome X, lies at the tail end of the null distribution, which may suggest an association between heterozygosity and these regions. HDR regions are characterized by repeated sequences, numerous mutations, and structural variants, complicating their mapping. Technical errors and paralogs/homologues mapped to the same location can also complicate their identification, increasing the probability that our heterozygous loci are mapping errors. Nevertheless, heterozygosity seems to be due to true biological variation between our populations and the reference genome. Chromosome X does not show an association between HDR and heterozygosity. It is important to note that this chromosome has the fewest SNPs (Figure S2) and HDR regions (Figure 5), making its framework slightly different from other chromosomes.

Using long-read sequencing to better characterize biological variation or changing the reference genome to suit our founder lines, using a pangenome reference, could avoid the mapping errors mentioned above and enable us to verify whether these are genuine heterozygotes.

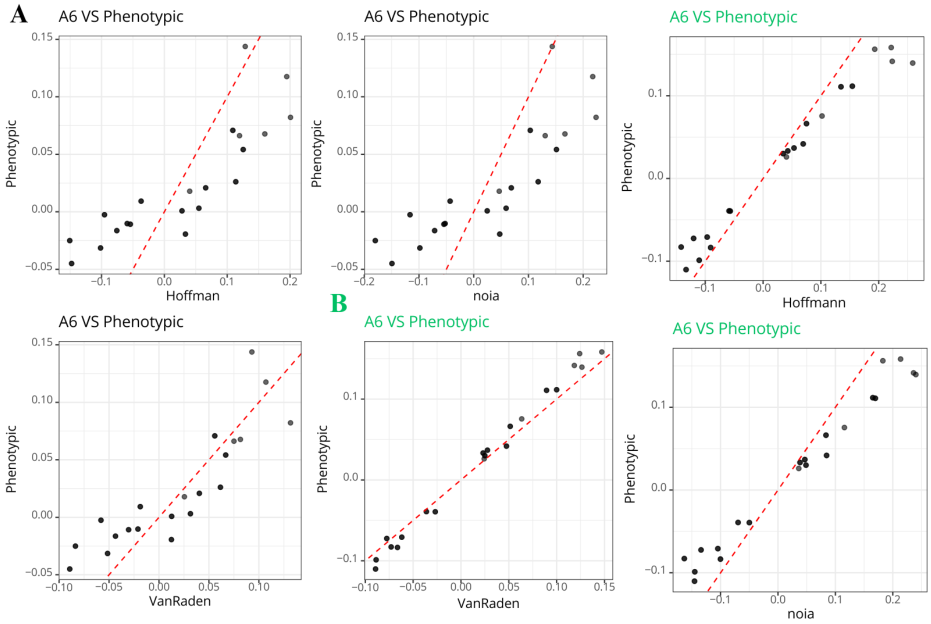

Figure 6: G-matrix comparison using different GRM constructions.
A: NGM condition **B**: NaCl condition

4.3 G-matrix

We compare our GBLUP model to a BLUP model using the G-matrix associated with each model to assess the impact of GRM definition. We compare each G-matrix to the BLUP one, named "Phenotypic" because it only takes into account phenotypic measures (Figure 6). In both conditions, the G-matrix across different GRM matrices is similar. Finding the same results with different methods shows that our data are good and our results robust. However, this means that for the considered population, its population structure, estimated by genetic similarity, is quite equivalent to considering all individuals as genetically independent from each other. This was discussed earlier in the GRM section, and this result was quite expected. If marker effects are known, then breeding values are exactly known because of their linear relationship. This result implies that if we obtain a new line from the same population and want to know its phenotype, we can sequence it. The SNP effects previously computed for the rest of the population allow us to calculate its breeding values as the sum of them (some environmental variables should also be taken into account). Thus, we should be able to predict the phenotype of a line based on its genotype. Due to the low population structure, we should be able to predict lines' phenotypes accurately.

VanRaden's method is the most consistent with results obtained using phenotypic data only, as in previous lab work, particularly under conditions of high salinity. Therefore, for the remainder of our analysis, we will select this method.

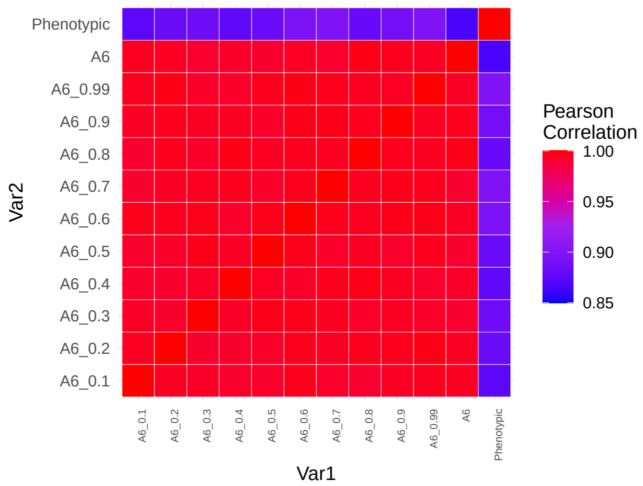

Figure 7: Heatmap depicting the correlation between G-matrices constructed from VanRaden GRM for different pruning thresholds and without GRM.

4.3.1 Pruning results

By retaining only certain SNPs, the structure of the kinship matrix, which measures genetic similarities between individuals, can be altered. Any modification of this matrix can influence the accuracy and reliability of the estimates. This alteration can affect the relative weights assigned to individuals in the GBLUP model, leading to variations in genetic predictions. Studying the impact of pruning on our results enables us to determine whether we can reasonably use pruned data.

The G-matrix obtained after different pruning steps was compared to confirm the equivalence between pruned and original data. We applied the multivariate GBLUP model for the A6140 population across all pruning thresholds. The comparisons, shown as Pearson correlation coefficients in Figure 7, range from 0.85 to 0.9, indicating high correlation and similar directionality. All matrices are highly correlated and similar to the G-matrix obtained without pruning (N). This high similarity across pruning stages suggests that pruning does not substantially affect the genetic variance-covariance structure, confirming that results with pruned data are equivalent to those with unpruned data.

4.4 Breeding values

4.4.1 Results for pruned data

To further verify the equivalence between a GWAS by GBLUP model with pruned or unpruned data, we plotted the mean and standard deviation of the medians of the posterior distributions of the breeding lines for each trait in multivariate analysis at different pruning thresholds, using the A6140 results in the NGM condition (Figure 8). The overlapping error bars and similar distributions across pruning thresholds indicate that pruning does not impact breeding values.

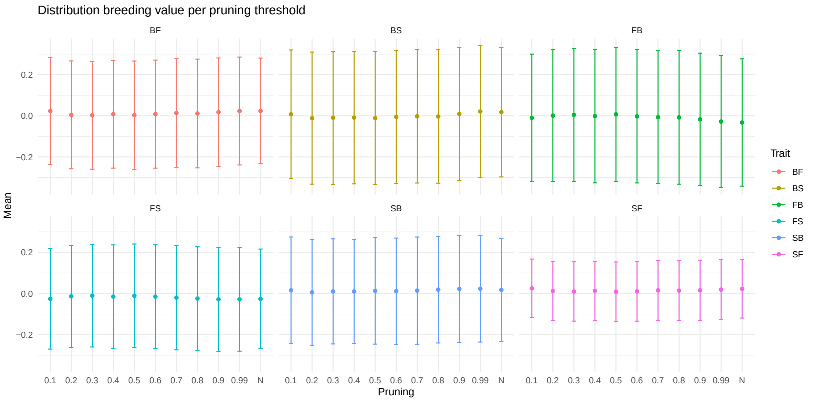

Figure 8: Mean of median Breeding values posteriori distribution among pruning

To reduce the number of SNPs considered, thus speeding up calculation times and simplifying result interpretation, we will use a pruning threshold of 0.99 for the rest of the analysis. This reduces the number of SNPs from 219,941 to 72,365. Since GBLUP works for both single and multiple traits, we consider our multivariate results are also valid in the univariate context.

4.4.2 Results for A6140 population

The difference between the multivariate and univariate approaches lies in the consideration of phenotypic covariance. The breeding value corresponds to the sum of additive effects in each line. We find that the standard deviation of the posterior intervals of breeding values has similar values but is shifted forward in the multivariate analysis (Figure S5). This indicates that the multivariate model captures higher variation, likely due to the covariance between traits.

Breeding values are fitted to a normal distribution with a mean of 0 and variance constrained by the GRM. With a 95% credibility interval, lines with breeding values statistically different from 0 can be identified (Figure S6). In the NGM condition, single-trait analysis reveals 38 lines with credible breeding values, 22 of which are associated with a single trait. In multi-trait analysis, 41 lines are identified as credible, with 18 lines credible for a single trait, meaning 56% are credible for at least two traits. This suggests a possible correlation between certain traits. Additionally, 2 lines are credible for all six traits in the multi-trait analysis, whereas none are identified for all six traits simultaneously in the single-trait analysis. This indicates that trait interactions are better captured in the multivariate analysis. Despite the limited number of credible lines per trait, the consistency of results indicates the model's robustness.

Interpreting these results in a Bayesian context is complex. Our prior assumed all lines belong to the same population, yet some lines show breeding values deviating from the mean. These deviations could result from errors, potentially from low-quality phenotypic data, though corrections for factors like temperature, density, session, and humidity should mitigate this. Alternatively, these could be true deviant phenotypes, indicating transgressive lines with extreme phenotypes due to

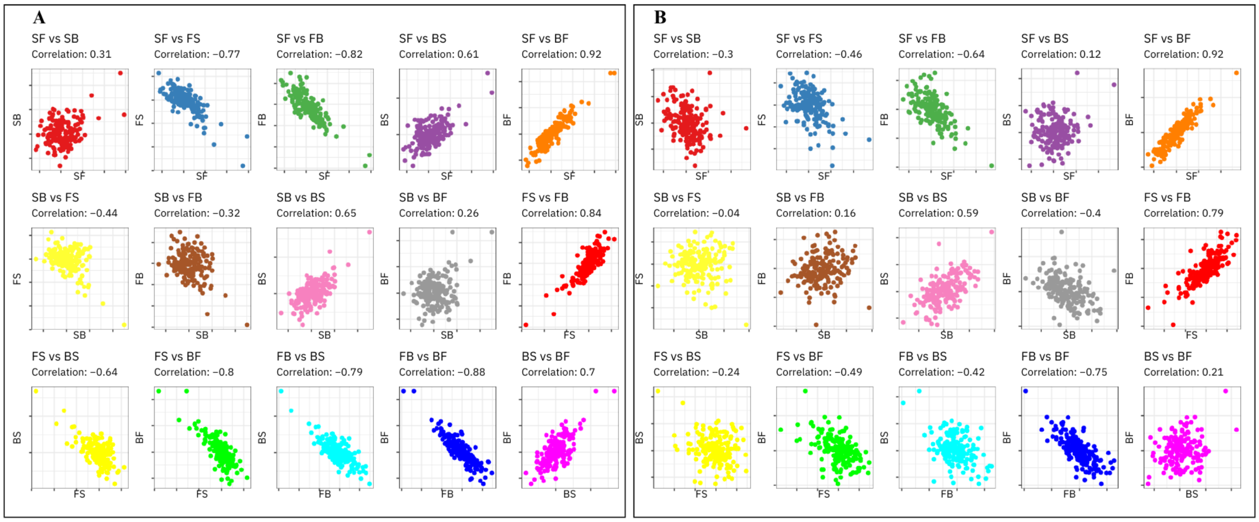

Figure 9: Correlation between traits for breeding values
A : NGM condition **B :** NaCl condition

epistasis and dominance. The NOIA matrix might better account for these lines by considering loci interactions.

One line, A6140L192, shows particularly deviant behavior in the multivariate analysis, with extreme credible values for all six traits. This line is not an isotype, eliminating potential bias. It is also associated in univariate analysis, but only for five traits. These lines appear to differ significantly from the population, potentially affecting our GWAS analysis. Excluding them and comparing results with and without these lines could clarify their impact.

To clarify the relationships between the traits, we plotted the medians of the traits compared two by two, where each point represents a lineage, using Pearson's coefficient to measure correlations. These results are depicted in Figures 9 **B** for multi-trait analysis and 9 **C** for single trait analysis. For the latter, correlation was observed only between the pairs of traits (SF-FB), (SF-BF), (SB-BS), (FS-FB), and (FB-BF). Whereas in multivariate analysis, all pairs of traits except SF-SB, SB-FS, SB-FB, and SB-BF are correlated. These results partly overlap with the correlations found by observing phenotypic values, although the SB-FS pair of traits was calculated as being correlated, which is not the case here. Breeding value corresponds to the additive genetic effect, so the results should align closely with those based solely on phenotypic measures. Since our lines are homozygous and the environmental component is accounted for by replicating measures, the phenotypic value can be considered equivalent to the genetic additive value. In multitrait analysis, as previously explained, environmental covariances are considered, making the correlations between traits similar to those found with phenotypic values. In contrast, univariate analysis does not account for covariances between traits, which can obscure correlations or result in correlations with the opposite sign compared to those calculated with phenotypic values. However, in this case, the signs of the correlations are mostly the same, suggesting that the environmental and genetic

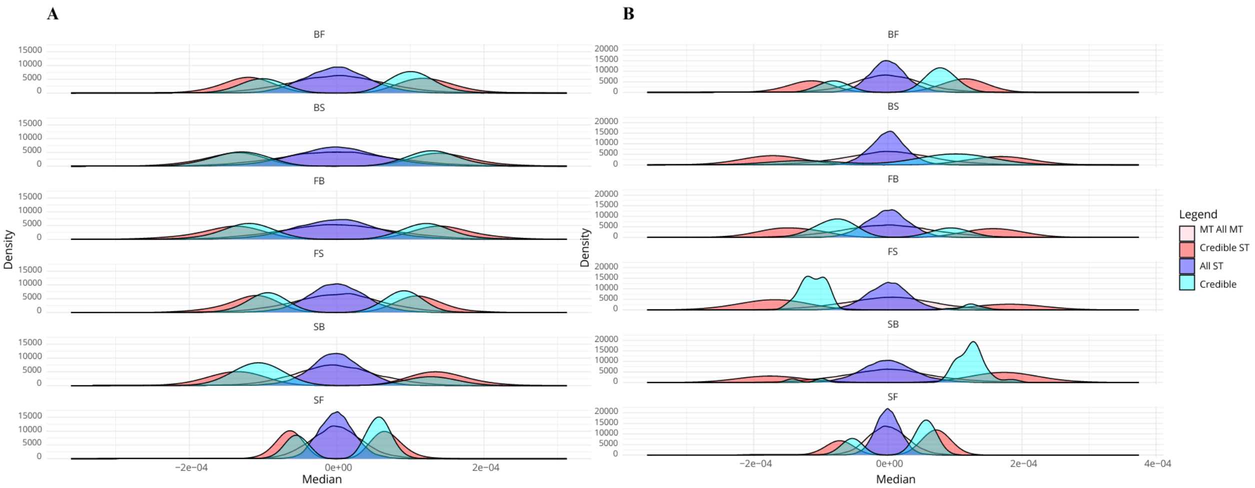

Figure 10: Density function of the median of the posterior distribution of SNP effects in single trait (ST) and multitrait (MT) analysis.
A: NGM condition **B**: NaCl condition

covariances likely share the same sign.

4.5 SNPs effects

Since backsolving for multiple traits is not widely accepted, we need to compare our univariate and multivariate analyses for consistency. We plotted the density of the medians of all SNPs and the credible SNPs for single-trait and six-transition-state analyses. Results under normal (Figure 10 **A**) and salinity (Figure 10 **B**) conditions, considered replicates, enhance the statistical robustness of our conclusions. In both conditions, the density function of all SNPs is a zero-centered Gaussian, which is consistent with the infinitesimal model where the majority of loci have small effects. As expected, the density function of the SNPs associated with the traits is bimodal, grouping together the extreme values of the SNP effects. In the NGM condition, the density curves show similarities between the different analyses, although the density of SNPs in the univariate analysis is slightly less extensive overall than in the multivariate analysis. The multivariate analysis seems to favor a wider distribution of SNPs. In the NaCl condition, the curves are more distinct, but the observations remain similar. The bivariate density curves for the effects of credible SNPs are more unbalanced in the single-trait condition than in the multi-trait condition, a trend already observed in the NGM condition but to a lesser extent. The density curve for all SNPs was symmetrical in both analyses and for both conditions, with the disparity possibly being due to the lower number of credible SNPs in the univariate analysis, which accentuated the differences. The GWAS multivariate analyses tended to reveal wider distributions of SNP effects than the univariate analyses. Two explanations can be proposed: first, there might be a bias in our model, which estimates higher effects in multi-trait analysis. The second hypothesis could be a biological explanation: multi-trait analysis takes into account the covariance between traits, which could extend estimates of SNP effects.

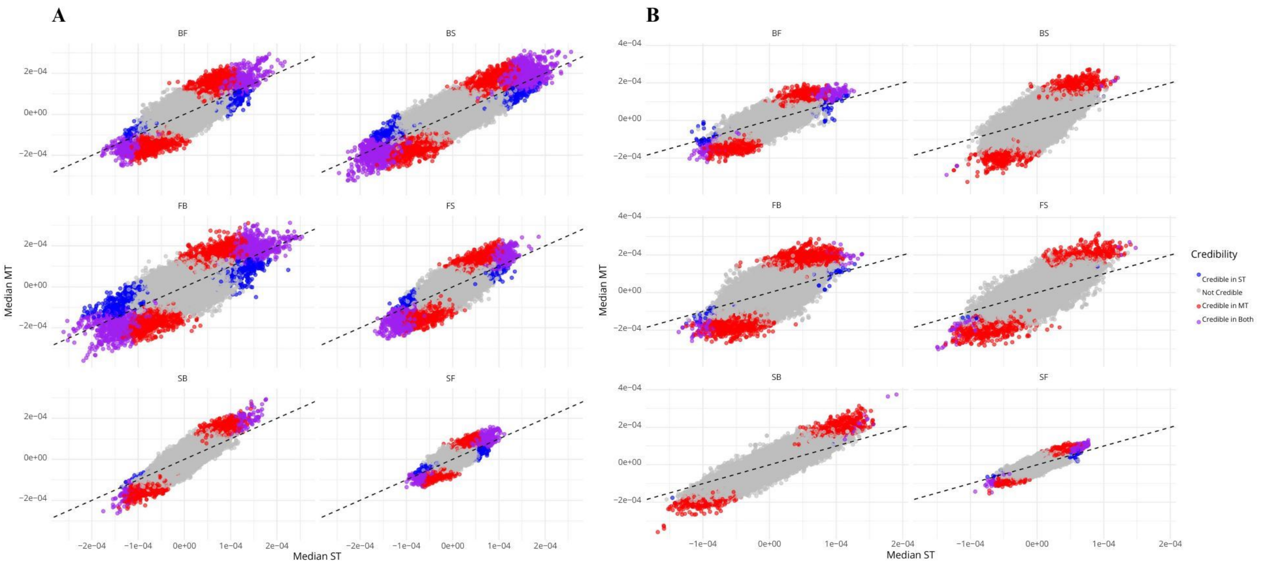

Figure 11: Median of the posterior distribution of SNP effects in multitrait (MT) analysis compared to single trait (ST) analysis.
A: NGM condition **B**: NaCl condition

However, to examine the impact of each type of analysis on SNP effects, we plotted the results (medians of the posterior distributions of SNP effects) of the multivariate analysis against the univariate analysis for both conditions (Figure 11).The first notable observation is the positive correlation between the univariate and multivariate analyses. In NGM conditions, the cloud of dots is tight, whereas they are more spread in the NaCl condition, indicating that the analyses are less correlated. In both environments, our observations have a slope steeper than the diagonal, meaning that SNP effects are higher in the multivariate analysis, which is consistent with our previous results. Although the effects of SNPs differ between the two analyses, they are highly correlated, which reinforces the validity of the multivariate approach.

To extend this analysis, bar plots were created for each condition, indicating the number of credible SNPs shared between the analyses, the number of SNPs present only in the multivariate analysis, and those present only in the univariate analysis. The latter are divided into two categories: those specific to one trait and those associated with several traits in the univariate analysis. Figure 12 **A** presents results for the NGM condition and Figure 12 **B** for the NaCl condition. These results provide a better understanding of the differences between the two types of analysis. It becomes clear that multivariate analysis identifies a greater number of SNPs associated with traits than univariate analysis: more than 50% of the SNPs associated with traits in the univariate analysis are also found in the multivariate analysis. These results, combined with the strong correlation observed between the results of the two analyses, lead us to conclude that multivariate analysis using backsolving provides satisfactory results. Multivariate analysis seems to offer a better ability to detect SNPs associated with the traits studied.

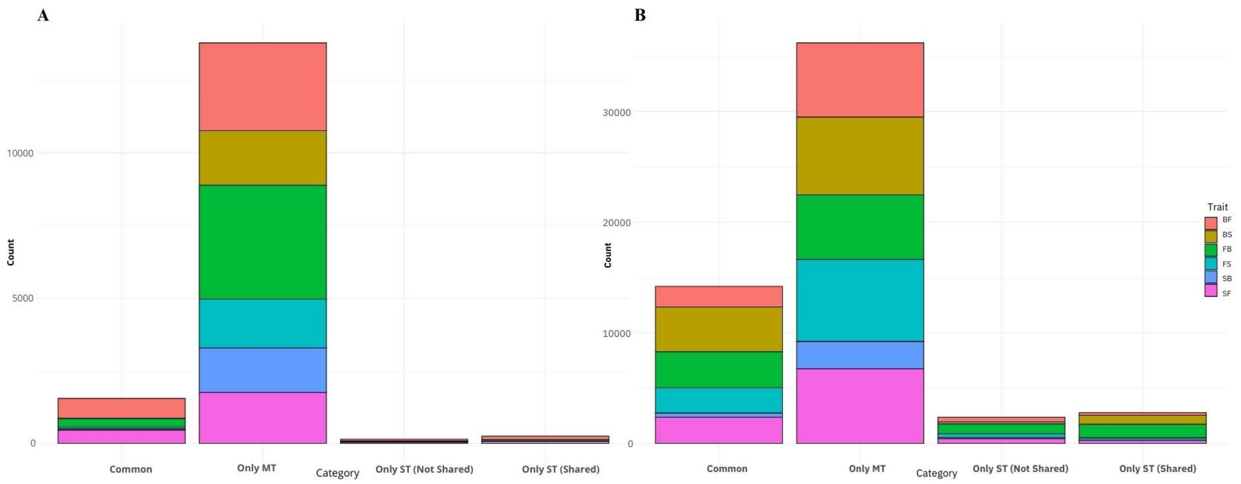

Figure 12: Repartition of SNPs with intervals not overlapping 0 for each trait between single trait (ST) analysis and multitrait (MT) analysis.
A: NGM condition**B**: NaCl condition

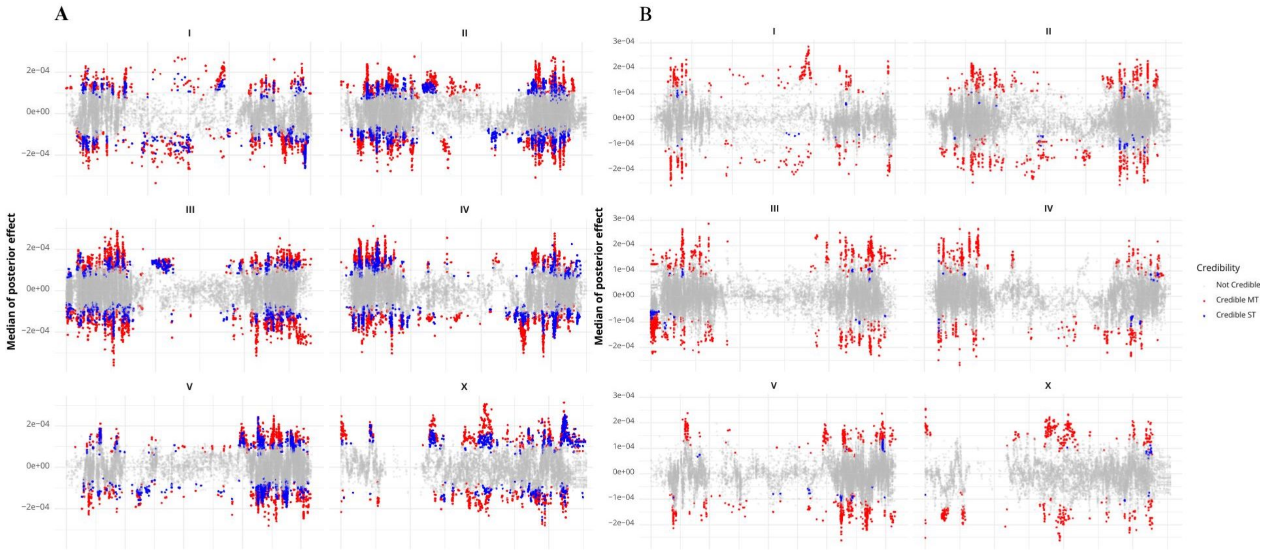

Figure 13: "Manhattan plot" depicting the median of the posterior distribution of SNP effects along the chromosome in single trait (ST) and multitrait (MT) analysis, focusing on FB.
A: NGM condition
B: NaCl condition

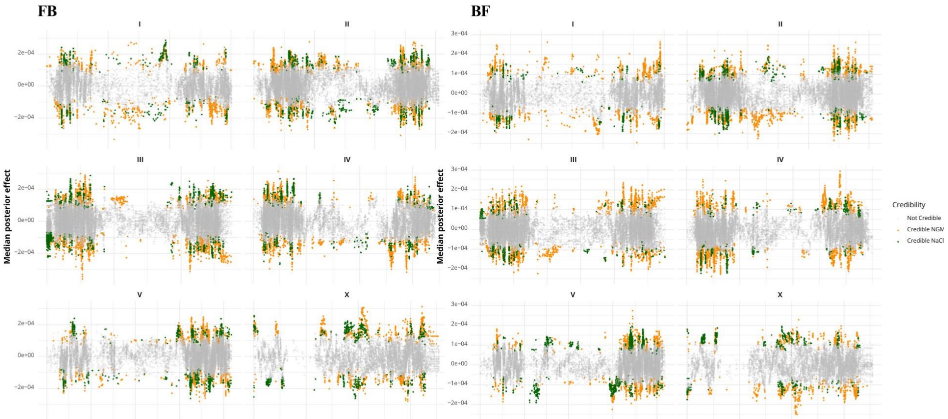

Figure 14: "Manhattan plot" depicting the median of the posterior distribution of SNP effects along the chromosome in normal and salt condition or FB and BF.

Now that we have established that the results of the multivariate analysis are consistent with those of the univariate analysis, we turn our attention to the identification of loci associated with transition states and locomotion in general. To identify the potential loci associated with transition states, we created Manhattan plots adapted to our a posteriori distributions. These plots represent the median of the a posteriori effects of the SNPs as a function of their positions on the chromosomes for normal (Figure 13 **A**) and salt (Figure 13 **B**) conditions. The red dots indicate the SNPs associated with the traits in the multivariate analysis, while the blue dots represent those identified in the univariate analysis. As the plots are similar for the different traits, we focus only on the Manhattan plots corresponding to the FB trait for a clearer presentation. This representation shows that the multivariate results enrich the univariate results, as the regions identified in the univariate analysis are found and extended in the multivariate analysis. Despite the pruning, there are still many signals. What is most noticeable is the distribution of credible SNPs on the arms of the chromosomes. Previous studies have shown that the loci influencing fitness are located at the ends of the chromosomes. *C. elegans* has holocentric chromosomes, with centromeres distributed along their length. Previous research in the laboratory has shown that recombination occurs mainly at the ends of chromosomes, increasing their potential for short-term adaptation through existing genetic variation and recombination. These results were expected since these traits have been shown to be under selection in previous studies [8] [9]. In the middle of chromosome III, we observed an unusual shape. Due to linkage disequilibrium, regions are expected to be in LD, and we would expect to see a chimney-like pattern. This anomaly could be an error due to our model or SNP calling. However, we find the same region in both analyses, suggesting that it is not linked to the backsolving step. It is also possible that the evolution in the middle of *C. elegans* chromosomes is unique and could result in this peculiar pattern.

To compare results between conditions, the same figures as above were plotted, with values in orange for the NGM conditions and green for the NaCl condition (Figure 14 **A**). The results pre-

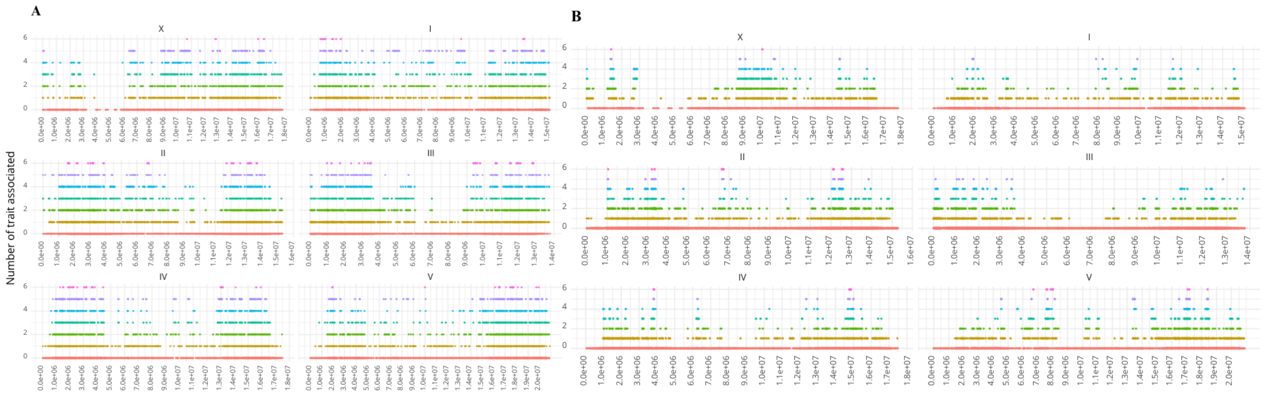

Figure 15: Repartition along chromosomes according to the number of traits with which they are associated **A :** NGM condition **B :** NaCl condition

sented are from the multivariate analysis. Almost all the regions identified in the salinity condition are also found in the normal condition, proving the robustness of our model and confirming our conclusions. These regions are associated with the FB transition in both environments, in the same direction and with very similar median effects. The loci identified appear to be robust candidates for further studies on the adaptation of the front-to-back transition in *C. elegans*. The BF transition was negatively correlated with breeding values (-0.88) and phenotypic data. The results were compared under both conditions in Figure 14 **B**. Several regions of the genome showed SNP effects inverse to those of FB: the right middle of chromosome 1, the left middle of chromosomes 2 and 3, and the beginning of the X chromosome.

Locomotion has been defined as being described by six transition states: SF, FS, BS, SB, FB, and BF. SNPs with an effect on these six traits can therefore be considered to influence overall locomotion. We have plotted the credible SNPs along the genome according to the number of traits with which they are associated in the multivariate analysis in normal (Figure 15 **A**) and salt (Figure 15 **B**) conditions. As expected, more regions associated with the six groups are present in the NGM condition. In both conditions, these regions are concentrated mainly at the ends of the chromosomes, which is consistent with previous observations that the ends of the chromosomes are zones of recombination and adaptation. These plots quantify pleiotropy, clearly showing that certain QTLs in the genome are associated with multiple traits. We observe that the distribution of regions associated with all traits varies between NGM and NaCl conditions. While our previous results suggested similar QTLs were found in both conditions, this is not confirmed here. For instance, although some QTLs associated with all traits appear in both conditions, such as one on chromosome two between 6e+06 and 7e+06, most QTLs identified under NaCl conditions are not found under NGM conditions.

4.6 Discussion

As before, one of the aims of genome-wide association studies is to understand the biological pathways between genotypes and phenotypes. Another use is to estimate the genetic propensities of individuals to predict their phenotype. Our observation of the G-matrix correlation between different GRM constructs leads us to the conclusion that we should be able to predict phenotype from genotype in our population using previous estimates of SNP effects. We can then consider using polygenic scores (PGS) that summarize the estimated effect of several SNPs on an individual's phenotype. Additionally, it has been shown that the joint use of several PGS, one PGS for each trait, allows for the combination of results from different genome-wide association studies, thereby increasing predictive power [22].

We find consistent results across different pruning thresholds, allowing us to conclude that the results obtained from the pruned data are similar to those obtained from the whole genome. We compared these results for the output of the GBLUP procedure, which showed that the breeding values were similar and so were the G-matrices. These findings support the use of the results from the backsolving step to reduce the number of SNPs in LD and clarify the interpretations. A previous study showed that applying linkage disequilibrium (LD)-based pruning to imputed whole genome sequencing data generally improved prediction accuracies for most traits. The study concluded that aggressive LD-based marker pruning significantly improves the accuracy of linear genomic unbiased prediction (GBLUP) when using imputed whole genome sequencing data, probably by reducing noise and redundancy in the dataset [23]. With these results, we can consider using only pruned data throughout our analyses, which would significantly speed up our results, given that the GBLUP step is the most computationally time-consuming.

We observed greater variances in breeding values in the multivariate model (Figure S5), suggesting that the model captures higher variation, likely due to the covariance between traits. Moreover, we also observed wider distributions of SNP effects in multivariate analyses compared to univariate analyses (Figure 10). Two explanations can be proposed: first, there may be a bias in our model, which estimates higher effects in multi-trait analysis, possibly due to shrinkage, but since the GRM is the same in both models, the shrinkage should be the same. The second hypothesis is a biological explanation: multi-trait analysis takes into account the covariance between traits, which could extend estimates of SNP effects. These two observations are coherent with each other: higher variance in breeding values leads to more extreme SNP effects.

We found many more variants in the multi-trait analysis, which is not surprising because several publications show that multivariate methods may offer a substantial increase in the discovery of genetic variants over the standard univariate approach [6] [24].

Our study suggests that SNPs identified for a specific trait in both analyses (Figure 12) are likely to be true positives. The SNPs present only in the multivariate analyses are associated thanks to the consideration of indirect selection and correlations, allowing a more precise, and in this case more extreme, estimate of their effects in line with biological reality.

For SNPs identified only in the univariate analysis, we distinguish two cases: on the one hand, SNPs considered as pleiotropic in the univariate analysis, where covariances between traits influenced the estimation of effects, leading to their overestimation in the univariate analysis. On the other hand, SNPs that are not pleiotropic in the univariate analysis may be associated with other traits in the multivariate analysis, in which case the inclusion of covariances has an impact on the distribution of SNP effects between traits.

In the case where non-pleiotropic SNPs identified in the univariate analysis are not associated with any trait in the multivariate analysis, this suggests that taking covariances into account has led to an underestimation of their effects in the multivariate analysis. However, it has been shown that, depending on the association scenarios between genotypes and phenotypes, univariate analysis may be more relevant than multivariate analysis, although the latter is generally better for understanding correlations between phenotypes. Consequently, some SNPs found only in the univariate analysis could be true positives and false positives in the multivariate analysis. It would be necessary to develop a method to validate these results and, if necessary, integrate these SNPs as QTLs. Additionally, some SNPs found only in one analysis could be linked to other SNPs and then considered as one QTL, but different SNPs representing it are flagged as associated in different analyses.

In addition to comparing univariate and multivariate methods, our results have enabled us to compare the effects of SNPs under normal and saline conditions (Figures 11, 12, 14, 15). These comparisons enable us to comment on the genetic architecture, i.e. the pattern of genetic effects that controls a given phenotype, under the two conditions. This architecture can be broken down into two parts: the number of QTLs involved in the expression of the trait and their distribution. For the first part, we observed that the number of QTLs was much lower in the NaCl condition. This stress condition may reveal cryptic genetic variations that have little or no effect under normal conditions but which manifest themselves under particular conditions, as in this case. Although a few new regions were identified in NaCl (Figure 14), the majority of the associated regions are similar to those found in NGM. The salt condition therefore reveals a few rare regions of cryptic genetic variation, but most of the regions identified in NaCl are the same as those in NGM, albeit less broad. This could be explained by the environment constraining certain QTLs, preventing them from expressing their effect. The large difference observed between the conditions could also be explained by the environmental variance, which is higher in the NaCl condition (Figure S7). Our model could therefore be subject to more noise, leading to under-detection of QTLs.

The second component is the distribution of these QTLs. As illustrated in Figure 15, the pleiotropic QTLs are not the same in the two conditions. This echoes a previous study in the laboratory, which showed that the evolution of a salt-adapted population could be predicted, but not its evolution in a normal environment [9]. The low overlap between SNPs obtained by comparing environments may explain - at least in part - why we can't explain the evolution of traits in the non-selected environment. Indeed, if most phenotype correlations between environments are explained by linkage (and not by pleiotropy), then these correlations evolve at the whim of recombination.

5 Conclusion and perspectives

Our study shows that the multivariate backsolving is consistent with the univariate results. By taking into account covariances between traits, we identified a greater number of SNPs. These results corroborate previous studies which suggest that multivariate GWAS detects more QTLs due to better statistical power.

To refine our analysis, several steps could be reconsidered. The GATK parameters, particularly those for filtration, could be optimized after simulating the variants, as proposed by the pipeline developed by CaeNDR, which includes simulated variants to choose thresholds adapted to our data. Additionally, our method for processing heterozygous loci could be reviewed. The filtration phase could be preceded by a step aimed at identifying whether heterozygotes are the result of sequencing errors and, if so, reassigning them as homozygotes. The results obtained with other genetic relationship matrices (GRMs), in particular NOIA, could be compared with the results presented here to assess whether this matrix better considers lineages deemed credible. It would also be interesting to repeat the GWAS, i.e., the backsolving stage, without the deviant lines that are likely to bias our results, and to compare these new results with those previously obtained to assess the impact of these lines.

Here, we have concentrated solely on the A6 population. It would be interesting to compare the results obtained with those of other populations. This would allow us to have a larger number of replicates to check the consistency of the multivariate analysis compared with the univariate analysis and to observe whether similar results are found in all populations. We expect to identify similar QTLs, as the other populations evolved from A6140, and the short-term adaptation we are considering is based mainly on pre-existing genetic variation.

To continue verifying the consistency of the results obtained in GWAS by multivariate GBLUP, it would be useful to carry out an analysis modeling the effects of the SNPs as fixed, using a SNP-by-SNP model, which has already proved effective in multivariate analysis. This approach would enable us to compare the SNPs identified as associated in the different analyses. If a coherent set of SNPs is found, this would further strengthen the validity of multivariate backsolving.

In GWAS, the problem of multiple testing arises when a large number of statistical tests are carried out simultaneously. This situation considerably increases the risk of false positives, where apparently significant genetic associations are in reality no more than statistical artifacts. Correcting for false positives is essential to ensure the reliability of the results. The False Discovery Rate (FDR) method of Benjamini and Hochberg [25] is commonly used to manage multiple tests in GWAS. However, this approach is not directly adapted to our Bayesian method, as it would require us to calculate the p-values associated with each SNP effect. It would be more interesting to use an empirical Bayesian approach to FDR, which uses effect sizes and their standard errors, as proposed in the method developed in [26].

Our study opens up a number of perspectives. Firstly, as mentioned in the discussion, phenotype prediction can be improved using polygenic scores (PGS). Additionally, by looking at the linkage between these QTLs in different populations, we could directly test the causality of recombination events on the evolution of genetic correlation between traits. More generally, there's a lot to explore about these QTLs and their evolution as a function of pleiotropy. The impact of these QTLs on adaptation could be measured with an adaptive architecture, which extends the notion of genetic architecture by including factors, such as the starting frequencies of the QTLs, that influence their adaptive potential [27]. This architecture takes the form of a probability distribution of the allele frequencies that contribute to the adaptive response. We could thus link quantitative genetics and population genetics, enabling us to explore alternative models of drift or selection, maybe finding characteristic signatures of polygenic adaptation as previously found [28].

Data and code availability Data, R code scripts, and modelling results can be found in the laboratory GitHub repository: `https://github.com/ExpEvolWormLab/Alix_internship`.

References

[1] Uffelmann, E., Huang, Q., Munung, N., De Vries, J., Okada, Y., Martin, A., Martin, H., Lappalainen, T. and Posthuma, D. (2021). Genome-wide association studies. Nat Rev Methods Primers *1*, 59.

[2] Sabatti, C. (2013). Multivariate Linear Models for GWAS p. 188–207. : Cambridge University Press.

[3] Clark, S. and Van Der Werf, J. (2013). Genomic Best Linear Unbiased Prediction (gBLUP) for the Estimation of Genomic Breeding Values, p. 321–330. : Humana Press.

[4] Legarra, A. (2018). Bases for Genomic Prediction.

[5] Legarra, A., Ricard, A. and Varona, L. (2018). GWAS by GBLUP: Single and Multimarker EMMAX and Bayes Factors, with an Example in Detection of a Major Gene for Horse Gait. G3 Genes|Genomes|Genetics *8*, 2301–2308.

[6] Stephens, M. and Balding, D. (2009). Bayesian statistical methods for genetic association studies. Nat Rev Genet *10*, 681–690.

[7] Wolc, A. and Dekkers, J. (2022). Application of Bayesian genomic prediction methods to genome-wide association analyses. Genet Sel Evol *54*, 31.

[8] Mallard, F., Noble, L., Guzella, T., Afonso, B., Baer, C. and Teotónio, H. (2023). Phenotypic stasis with genetic divergence. Peer Community Journal *3*, e119.

[9] Mallard, F., Afonso, B. and Teotónio, H. (2023). Selection and the direction of phenotypic evolution. eLife *12*, e80993.

[10] Stephens, M. (2013). A Unified Framework for Association Analysis with Multiple Related Phenotypes. PLoS ONE *8*, e65245.

[11] Noble, L., Chelo, I., Guzella, T., Afonso, B., Riccardi, D., Ammerman, P., Dayarian, A., Carvalho, S., Crist, A. and Pino-Querido, A. (2017). Polygenicity and Epistasis Underlie Fitness-Proximal Traits in the Caenorhabditis elegans Multiparental Experimental Evolution (CeMEE) Panel. Genetics *207*, 1663–1685.

[12] Noble, L., Rockman, M. and Teotónio, H. (2021). Gene-level quantitative trait mapping in Caenorhabditis elegans. G3 Genes|Genomes|Genetics *11*, jkaa061.

[13] Teotónio, H., Estes, S., Phillips, P. and Baer, C. (2017). Experimental Evolution with Caenorhabditis Nematodes. Genetics *206*, 691–716.

[14] Parée, T. (2023). Origin and Evolution of Recombination Rate Landscapes Diversity An Experimental Evolution approach. PhD thesis, Univeristé PSL.

[15] Lee, D., Zdraljevic, S., Stevens, L., Wang, Y., Tanny, R., Crombie, T., Cook, D., Webster, A., Chirakar, R., Baugh, L. et al. (2021). Balancing selection maintains hyper-divergent haplotypes in Caenorhabditis elegans. Nature Ecology & Evolution *5*, 794–807.

[16] VanRaden, P. (2008). Efficient Methods to Compute Genomic Predictions. Journal of Dairy Science *91*, 4414–4423.

[17] Joshi, R., Meuwissen, T., Woolliams, J. and Gjøen, H. (2020). Genomic dissection of maternal, additive and non-additive genetic effects for growth and carcass traits in Nile tilapia. Genet Sel Evol *52*, 1.

[18] Hoffman, G. (2013). Correcting for Population Structure and Kinship Using the Linear Mixed Model: Theory and Extensions. PLoS ONE *8*, e75707.

[19] Gualdrón Duarte, J., Cantet, R., Bates, R., Ernst, C., Raney, N. and Steibel, J. (2014). Rapid screening for phenotype-genotype associations by linear transformations of genomic evaluations. BMC Bioinformatics *15*, 246.

[20] Wang, H., Misztal, I., Aguilar, I., Legarra, A. and Muir, W. (2012). Genome-wide association mapping including phenotypes from relatives without genotypes. Genet. Res. *94*, 73–83.

[21] Verardo, L., Silva, F., Lopes, M., Madsen, O., Bastiaansen, J., Knol, E., Kelly, M., Varona, L., Lopes, P. and Guimarães, S. (2016). Revealing new candidate genes for reproductive traits in pigs: combining Bayesian GWAS and functional pathways. Genet Sel Evol *48*, 9.

[22] Krapohl, E., Patel, H., Newhouse, S., Curtis, C., Von Stumm, S., Dale, P., Zabaneh, D., Breen, G., O'Reilly, P. and Plomin, R. (2018). Multi-polygenic score approach to trait prediction. Mol Psychiatry *23*, 1368–1374.

[23] Ye, S., Gao, N., Zheng, R., Chen, Z., Teng, J., Yuan, X., Zhang, H., Chen, Z., Zhang, X. and Li, J. (2019). Strategies for Obtaining and Pruning Imputed Whole-Genome Sequence Data for Genomic Prediction. Front. Genet. *10*, 673.

[24] Porter, H. and O'Reilly, P. (2017). Multivariate simulation framework reveals performance of multi-trait GWAS methods. Sci Rep *7*, 38837.

[25] Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B: Statistical Methodology *57*, 289–300.

[26] Stephens, M. (2016). False discovery rates: a new deal. Biostat *kxw041*.

[27] Barghi, N., Hermisson, J. and Schlötterer, C. (2020). Polygenic adaptation: a unifying framework to understand positive selection. Nat Rev Genet *21*, 769–781.

[28] Franssen, S., Kofler, R. and Schlötterer, C. (2017). Uncovering the genetic signature of quantitative trait evolution with replicated time series data. Heredity *118*, 42–51.

Appendix

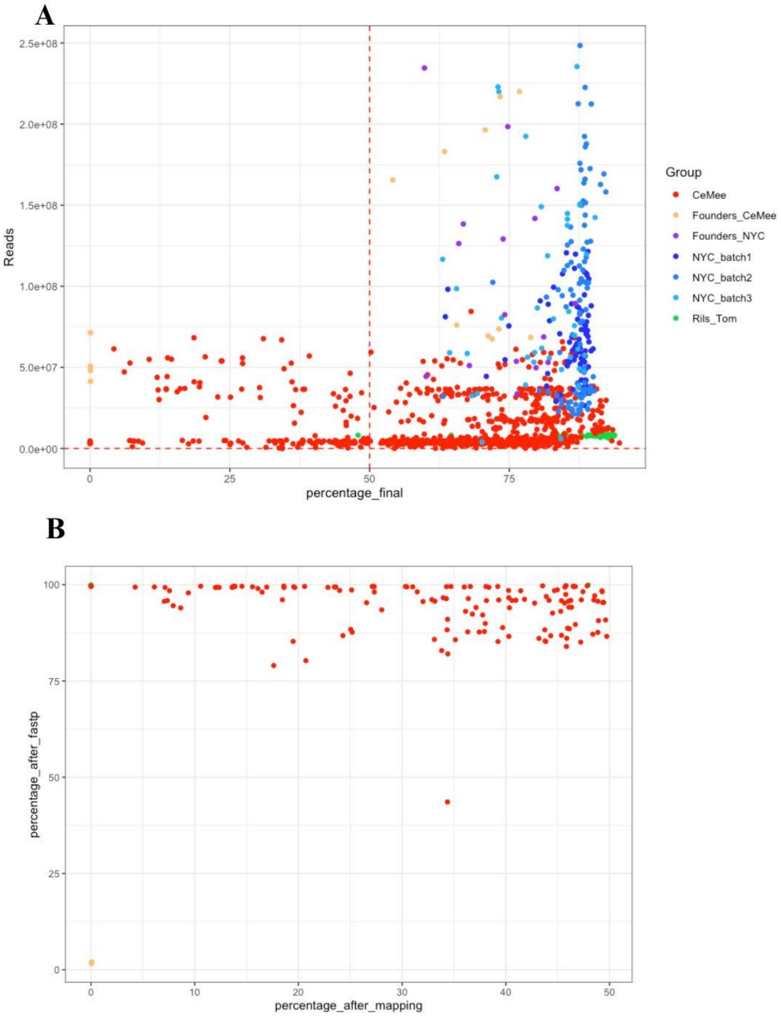

Figure S1:

A: Number of reads sequenced against number of reads after mapping filtering and filtering by `fastp`. This tool removes low-quality reads and performs trimming (i.e., removes Illumina adaptors) and filtration after mapping to the reference genome WS283. The correlation between the number of reads and coverage is as expected, based on the theoretical formula for coverage: $\frac{\text{Number\_of\_reads} \times \text{Average\_read\_length}}{\text{Size\_of\_the\_genome}}$.

B: Percentage of reads retained after the `fastp` and mapping steps. This figure highlights the origin of read loss, showing that five founder lines from the CeMEE sequencing lost almost all their reads after the `fastp` step, indicating poor sequencing quality.

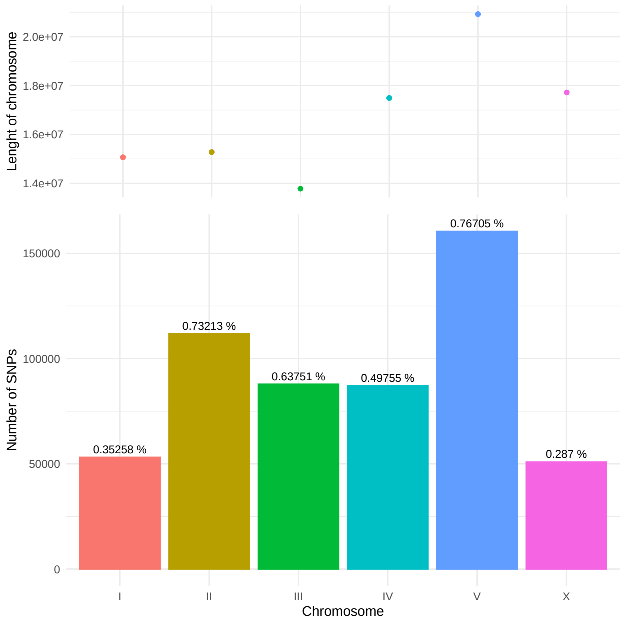

Figure S2:

Top:Chromosome length

Bottom: Number of SNPs per chromosome, proportion $\frac{\text{Number SNPs}}{\text{Lenght chromosome}}$ on top of bars. Chromosome lenght is not proportionnal to the Number of SNPs.The X chromosome appears to be the most conserved between populations, while chromosome 5 shows greater diversity.

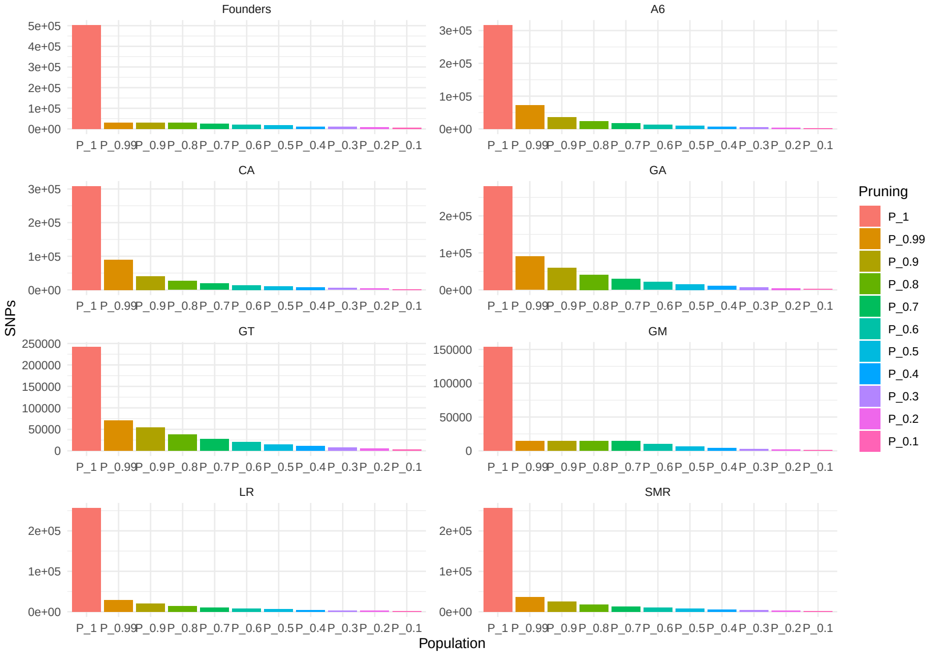

Figure S3: Pruning Results All Population

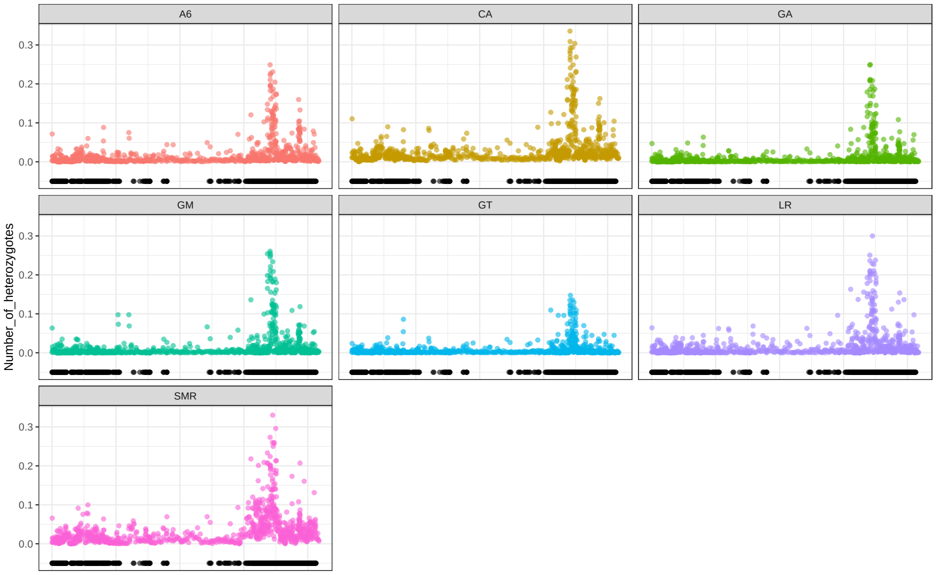

Figure S4: Heterozygotie of all population along chromosome V

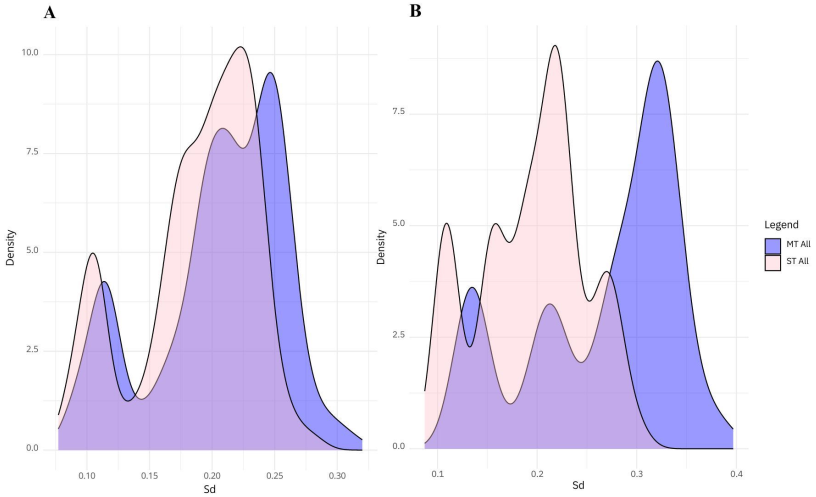

Figure S5: Density function of the standard deviation of the breeding value posterior distribution.

A: NGM condition **B**: NaCl condition

The x-axis depicts the standard deviation of each SNP effect distribution a posteriori, while the y-axis represents the probability density, not the probability itself. This value is not constrained between 0 and 1 because it reflects how dense the data points are in a given region rather than the likelihood of a specific outcome. The key property of the density plot is that the total area under the curve equals 1, representing the entirety of the probability distribution. This means that the y-values can be greater than 1 as long as the integral of the density over the entire range of the data is equal to 1.

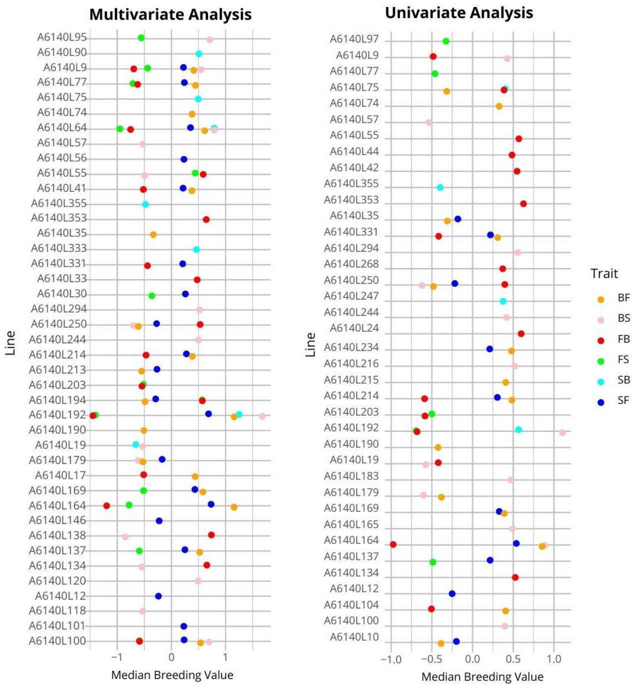

Figure S6: Forest plots comparing lines with breeding values whose credibility intervals do not overlap 0 in univariate and multivariate analyses.

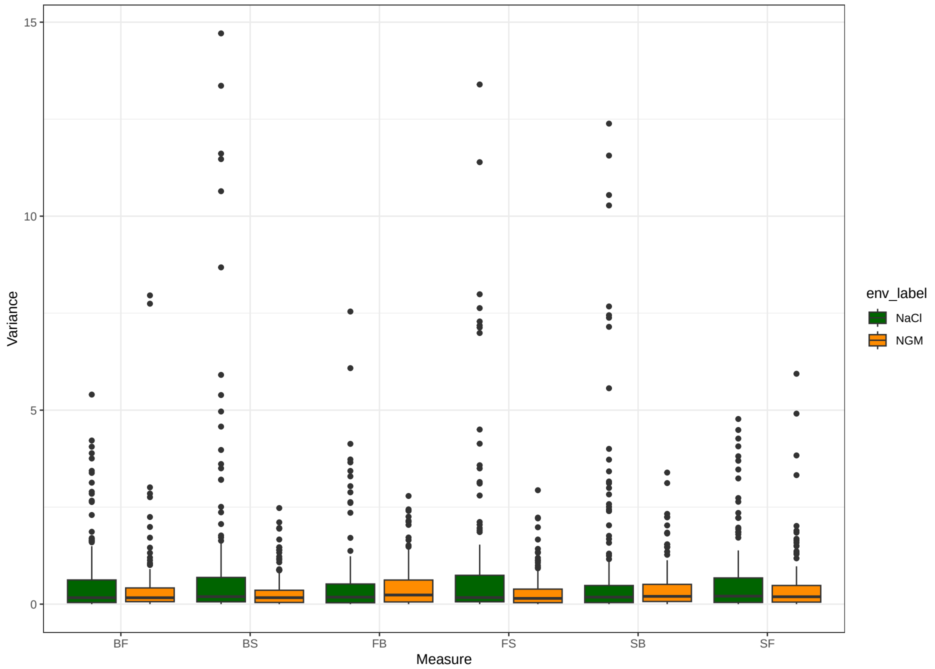

Figure S7: Environmental variance measured for each trait across conditions. NaCl shows greater variation in four traits (BS, FB, FS, SB). The same lines (A6140L247 and A6140L103) exhibit high variance in BF and SF. A6140L103 belongs to an isotype group, whereas A6140L247 does not, indicating that this high variance is not biased by group membership. Despite the high variance in these two lines, environmental variance is generally higher in the NaCl condition.