Title: Strength of stabilizing selection on the amino-acid sequence is associated with the amount of non-additive variance in gene expression

Authors: Margarita Takou[1,2], Daniel J. Balick[3], Kim A. Steige[1,4], Hannes Dittberner[1], Ulrike Göbel[1,5], Holger Schielzeth[6], Juliette de Meaux[1]

Addresses:

1 Institute of Botany, University of Cologne, Cologne, Germany

2 Department of Biology, Pennsylvania State University, PA, United States of America

3 Department of Biomedical Informatics, Harvard Medical School, Boston, Massachusetts, United States of America

4 State Plant Breeding Institute, University of Hohenheim, Stuttgart, Germany

5 Bioinformatics Core Facility, CECAD Research Center, Cologne, Germany

6. Institute of Ecology and Evolution, Friedrich Schiller University Jena, Jena, Germany

Abstract:

The distinction between the additive and non-additive components of genetic trait variance has long been known to be crucial for predicting the potential of a population to respond to selection. Yet we know surprisingly little about the genomic and evolutionary factors that influence non-additive genetic variance in natural populations. Here, we use a quantitative genetic breeding design to delineate the additive and non-additive components of expression variance of 17,657 gene transcripts in the outcrossing plant *Arabidopsis lyrata*. We find that most of the genetic variance in gene expression is from non-additive variance. Transcripts showing the highest levels of non-additive variance tend to be longer, clustered in larger groups of co-expressed genes, and enriched in functions related to cellular differentiation or to the epigenetic reprogramming of gene expression. Furthermore, we find that genes with the most non-additive variation not only display a markedly lower rate of synonymous variation, they are also exposed to stronger stabilizing selective forces than genes with high additive variance. To the best of our knowledge, this study is the first to show that stabilizing selection on amino-acid sequences is associated with a decrease in a population's potential for adaptation of gene expression levels.

Main

Genetic variation is a prerequisite for populations to respond to selection[1]. However, in each generation, alleles within the same or different loci are reshuffled, so that the phenotypic effects resulting from inter-allelic interactions cannot be transmitted from parents to offspring. Consequently, only a fraction of the genetic variation in a trait exposed to selection can sustain a response to selection. To account for this fraction, Fisher divided genetic variance into additive variance, to refer to components that are directly transmitted to the next generation, and non-additive variance, to describe the remaining variance[2]. The main non-additive component is termed dominance variance (V_D) in quantitative genetics studies, and although strictly defined for intra-locus interactions, the use of the term typically also includes inter-locus allelic interactions that give rise to epistatic genetic variance. The amount of non-additive variance inferred in quantitative genetics studies is a statistical component of variation. It cannot be simply derived from the recessive/dominant relationships between alleles that can be dissected genetically[3]. Because most classic genome-wide association studies (GWAS) assume co-dominant allelic relationships, the amount of trait heritability in a population that has the potential to respond to selection remains unknown[4–9] (but see[3,10]). This amount also cannot be derived from dissecting interactions between alleles at different loci because a fraction of functional epistasis will contribute to additive variance[3,11,12]. Therefore, quantitative genetics approaches are required to understand what controls how much genetic variance present in a natural population is non-additive and thus not available for selection.

Several factors are expected to influence the amount of non-additive genetic variance. First, the amount depends on the genetic architecture of the phenotype. As the number of loci controlling a phenotype increases, so will the likelihood that new allelic variants interact with other variants of the same or other loci to define a phenotypic value that differs from the sum of effects of the individual variants[13]. Second, the amount of non-additive variance is assumed to depend on the population's recent history. Indeed, non-additive variance is expected to be enriched in traits that have been depleted of their additive genetic variance by recent natural selection[14,15]. Indeed, additive variance will decrease after the advantageous alleles have fixed in the population[7]. By contrast, when genetic drift is the primary force affecting allele frequencies, the proportion of

additive genetic variance is expected to increase[7,16,17]. The amount of non-additive variance thus depends on the molecular architecture of a trait as well as on the selective and demographic history of the population.

Despite the crucial importance of non-additive variance for predicting the adaptive potential of single traits, few studies have investigated the evolutionary factors underlying the amount of non-additive variance present in natural populations, and even less is known about the molecular and genomic features that promote the emergence of non-additive genetic variation[16]. Transcriptome studies generate powerful opportunities to study the evolutionary properties of a large collection of quantitative molecular traits[18]. In addition, such approaches can deliver crucial information on the extent to which genetic variation in gene expression can contribute to adaptation[19–21]. Most transcriptome studies assume that genetic variance is additive[18,21]. However, when quantified over appropriate pedigrees, transcriptome variation provides ideal data for exploring the genomic and evolutionary properties shaping the amount of non-additive variance in natural populations. Indeed, transcript level variance in full- and half-sib families of threespine sticklebacks showed that non-additive variance is pervasive in the transcriptome[22]. Conducting such studies in plants can be particularly powerful because environmental and maternal variance can be easily distinguished, preventing them from having a confounding effect. Here, we quantify the additive and non-additive components of genetic variance in gene expression during the exponential growth phase of the outcrossing plant species *Arabidopsis lyrata*. We show that additive and non-additive variance in gene expression varies with the level of negative selection detected at the protein level. Whereas additive variance tends to be associated with genes experiencing relaxed selection, we find that non-additive variance is pervasive, dependent on transcript length, and promoted by the complexity of gene regulation.

Results

We took advantage of a set of 10 full- and half-sib families of the outcrossing plant species *Arabidopsis lyrata* ssp. *petraea*, which were obtained by crossing individuals from two locally adapted populations [23–25] (Supplementary Table S1). Given the amount of shared variation between the populations and their similar frequency distribution, this set of inter-population hybrids is expected to show levels of non-additive variance similar to those within populations (Supplementary Figure S1)[26]. Full- and half-sib relationships allowed us to partition the total

phenotypic variance (V_P) of gene expression into additive variance (V_A), non-additive (V_D), maternal (V_M), and residual (V_R) components of variance in the expression for each transcript in the genome (Figure 1a, Supplementary Table 1). We refer to V_D here as the non-additive variance because the statistical variance due intra- and inter-locus, i.e. dominance and epistatic variance, cannot be disentangled. In 67.7% of the 17,657 genes analyzed, more than 50% of total expression variance (V_P) was explained by the significant contribution of the two genetic components of variance, V_D (mean = 0.375) and V_A (mean = 0.215). The maternal variance (V_M) was close to zero for most transcripts (mean = 0.088). Finally, the population of origin of the mother plant did not affect gene expression variance (Bonferroni corrected threshold of p = 2.486e-06).

Delineating overall genetic variance showed that V_D contributed the most (25.4% of 17,657 genes had more than 50% of gene expression variance V_P due to V_D), in contrast to the additive genetic component V_A (6.3% of these genes had more than 50% of their variance in expression, Figure 1b-c). Hence, gene expression variance V_P is due largely to genetics, but only a limited fraction of that genetic variance is inherited additively. Although amounts of non-additive genetic variance are rarely quantified in natural populations, the amount we detected is not unexpected. Similar fractions of non-additive variance have been reported in phenotypic traits[16], and non-additive variance was as high if not higher than additive variance for most transcripts expressed in threespine stickleback[22]. In a control dataset generated by permuting family labels, V_R explained more than half of the V_P for more than 73% of transcripts (mean = 0.717), and V_D was not significant (mean = 0.135), confirming that the levels of genetic variance we report are not random.

Most genetic variance cannot be explained by large effect regulatory variants

We used our data to investigate the genomic and evolutionary features that are associated with the relative importance of additive and non-additive variance in gene expression. We first explored the regulatory architecture of gene expression, using a hierarchical clustering approach. We used a k-clustering method and tested a number of clusters k ranging from 23 to 300 clusters of co-expressed genes (Supplementary Figure S2). Setting the number of clusters, k, to 200, cluster size ranged between 5 and 2,829 genes (median=39.5 genes per cluster). Thus, and

perhaps unsurprisingly given what is known about the regulatory networks controlling gene expression[18,27], our dataset revealed that many genes form co-expressed clusters of variable size.

To evaluate the number and effect sizes of regulatory variants explaining the expression and co-expression of the genes, we used allelic segregation across 131 individuals of the transcriptome sample; this allowed us to implement a GWAS approach, controlling for the family structure of the mapping population (see Methods). In total, we identified 156 Single Nucleotide Polymorphisms (SNPs) associated with the expression level of only 174 of the 17,656 expressed gene transcripts and these 156 SNPs never co-localized with the transcript (Supplementary Figure S2). Thus, for 99% of expressed transcripts, no main causal genetic variant could be detected with the approach. Importantly, most of the 156 SNPs (76.4%) were associated with the expression of a single gene and, thus, do not explain the formation of any of the clusters. This result suggests that most of the genetic variation in expression has a polygenic basis. A major genetic incompatibility, which can lead to phenotypic aberration and broad regulatory alterations in inter-population crosses, is unlikely to explain the data[28–31].

Differences in the strength of negative selection are associated with the mode of regulatory inheritance

Both drift and natural selection are predicted to shape the amount of non-additive variance in a population [11,12]. To determine whether the non-additive variance in gene expression is linked to the evolutionary forces acting on the encoded protein, we examined rates of amino-acid evolution in the transcripts that were expressed in our dataset. We partitioned expressed transcripts in three non-overlapping sets based on their relative excess of V_D, V_A, or V_R (Figure 2a). As a control, we included a fourth set of 3,500 genes sampled at random (Fig. 2a). We then estimated the rate of amino acid divergence that evolved between *A. lyrata* and *A. thaliana* across these four groups of genes. The four sets of transcripts differed significantly in their rate of amino-acid divergence. Genes with high V_R had the lowest rate of amino-acid divergence, whereas genes with highest V_A had the fastest rate (Figure 2b-c). Interestingly, genes with the largest V_D displayed a rate of amino-acid divergence that was intermediate between those shown by genes with large V_A and genes with large V_R.

The rate of amino-acid divergence is expected to increase if positive selection is acting on the encoded protein, but it can also arise if the evolutionary constraints on the amino-acid sequence

are weak. In order to differentiate these alternatives, the segregation of molecular variation within species can be used to infer the fitness effect of new mutations and quantify the strength of the evolutionary constraints imposed on the amino-acid sequence. We took advantage of a previous study, which determined the demographic history of the same two parental populations of *A. lyrata*[25,32]. We inferred the distribution of fitness effects (DFE) of new non-synonymous variants in these gene sets, following a method described in[25] that incorporates the specific demographic parameters of each population (Supplementary Figure S3). Both the inferred demographic model and the DFE fitted the site-frequency spectra (SFS) of synonymous and non-synonymous variants well (Supplementary Figure S3a), indicating that the model provides a good representation of the evolutionary and selective history of these populations.

The results showed that the inferred strength of evolutionary constraint (Figure 2d; Supplementary Table S2) co-varies with differences in the mode of genetic variance in transcript expression level. Compared to the other sets of genes, the set of genes with low genetic variance in expression (i.e., high residual variance, V_R) showed a significantly lower proportion of nearly neutral non-synonymous mutations ($0 < N_{anc}s < 1$, $p = 0.01$), as well as a significantly higher proportion of mutations under strong purifying selection ($N_{anc}s > 10$, $p = 0.01$). This observation indicates that genes evolving under strong constraint at the amino-acid level also tended to show low levels of genetic variation at the transcript expression level. Stabilizing selection has indeed been shown to be associated with decreased rates of expression variation or divergence [18,19,33–39].

Our analysis further revealed that selective constraints on genes with large V_A (threshold 0.5) are comparatively lower than on other genes (Figure 2c, Supplementary Figure S3b-c). The DFE of the set of genes with large V_A in transcript expression indicated that, compared to other gene sets, these genes hosted a significantly larger fraction of new non-synonymous mutations that are nearly neutral ($0 < N_{anc}s < 1$, $p= 0.03$). At the same time, a significantly smaller fraction of new mutations is under strong purifying selection in genes with large V_A ($N_{anc}s > 10$, $p = 0.01$). Furthermore, genes with high V_A were less likely to map to genomic regions harboring footprints of selective sweeps. Although regions with signatures of selective sweeps extended over 27% of the genome, only 17% of the high V_A genes mapped to these regions (pairwise t-test, $p= 2.115e-12$), further confirming that these genes are unlikely to contain variants that have been under directional selection. Therefore, both the high rate of amino-acid divergence observed in these

genes and their inferred DFE indicate that genes with additive variance in gene expression tend to be exposed to lower levels of purifying selection than are the other genes.

The evolutionary constraints detected on the amino-acid sequence of genes with high levels of V_D (threshold 0.5) did not differ significantly from those detected in the control group of random genes (Fig. 2c, Supplementary Figure S3d). However, the rate of nucleotide divergence at synonymous sites, Ks, was markedly lower in genes with high V_D, than in all other gene groups (Supplementary Figure S3). A lower K_S is expected if the mutation rate is lower or the background selection is stronger[40].

Gene architecture associates with the level of non-additive variance in transcripts

In order to better understand what modulates the relative fractions of non-additive and additive variance, we compiled a total of 200 descriptors of genetic variation at each gene, including the number of known transcription-factor binding sites in the upstream non-coding region, the length of the genes, and various descriptors of nucleotide diversity (see Methods). We trained a random forest model to identify the genomic architecture or population genetic parameters that best explain variation in the prevalence of V_A or V_D in gene expression. We first performed the analysis using the 200 descriptors to predict variation in V_A, but the model's ability to classify high V_A genes was poor (R^2 = 0.009 and mean square error of 0.021).

For variation in V_D, however, the model was accurate at classifying genes with high V_D (R^2 = 0.18 and mean square error of 0.026), highlighting that the most important factors for classifying genes according to their level of V_D are descriptors of gene structure (Figure 3a; Supplementary Table S3). Specifically, the number of exons per gene (importance: 53.021; p = 0.0164), total gene length (importance: 26.92; p = 0.0164), and transcript length (importance: 14.27; p = 0.0164) were the three factors that contributed most to predict levels of non-additive variance. The importance of exon number persisted even when gene length was corrected for. Transcript length was indeed significantly correlated with V_D (Spearman coefficient ρ = 0.258, p = 1.22e-266, Supplementary Table S4; Figure 3b). We used a linear model to confirm that the effect of transcript length on V_D (p < 2.2e-16) remained significant even after correcting for median gene expression (p = 0.0176) and V_G (p < 2.2e-16). The model did not suggest a significant interaction between transcript length and median gene expression (p = 0.224). The number of exons per gene correlated significantly with the level of non-additive variance (ρ = 0.358, p < 2.2e-16;

Figure 3c). The random forest analysis further revealed that the density of SNPs within a gene was a predictor of high V_D. Strikingly, all significant descriptors tend to increase the mutational target size of the transcript. We thus conclude that non-additive variance is more likely to be observed as the mutational target size increases.

Interestingly, the analysis further revealed that the total number of transcription-factor binding sites explained V_D well (p = 0.0163; importance 4.34; Figure 3d). When we corrected for gene length, the total number of transcription-factor binding sites becomes the most important predictor in the random forest analysis (Importance 28.72; p =0.0163), which explains the level of V_D in the dataset. In particular, the transcription factor binding sites associated with high V_D were targets of a DOF zinc finger protein expressed in young rosettes (DOF 5.3; p = 0.0163), that was recently shown to determine stem cell division necessary for plant thickening[41] (AT5G60200 TMO6; p = 0.01639). It is likely that this regulator was active in the tissue samples we collected, because we sampled during the exponential phase of rosette growth.

We further quantified the distribution of genes with high V_D across Gene Ontology (GO) categories of expressed genes. We found that genes with high V_D were enriched for GO categories reflecting cell differentiation. The top enriched GO terms for the highest V_D estimates were "chromatin silencing" (p = 4.5e-16) and "cytokinesis by cell plate formation" (p = 2e-11), "DNA methylation" (p = 4.3e-11; Supplementary Table S5). Genes with high V_A were also not randomly distributed among molecular functions, but, in contrast to V_D genes, their functional enrichment was less striking than for high V_D genes and not related to cell differentiation. Among the top GO categories most enriched among genes with high V_A values were several functions linked to central metabolism such as "glucose catabolic process" (p = 8.3-05) and "glycolytic process" (p = 1.6e-05), but we also detected enrichments among genes involved in the "response to cadmium ion" (p = 0.00015) and "proteasome core complex assembly" (p = 0.00016; Supplementary Table S6).

Non-additive variance increases with the size of the co-regulatory network

Our results indicate that two processes may increase the opportunities for the inter-allelic interactions that favor the emergence of non-additive variance: the mutational target size of a transcript (its length and number of exon-intron boundaries) and the complexity of the regulatory network. The observations that transcription-factor binding sites as well as the DOF zinc finger

protein TMO6 contribute significantly to the amount of non-additive variance in gene expression confirm this observation. More generally, we found that the median V_D in a cluster of co-regulated genes increased with the number of genes contained in a cluster (ρ = 0.49, p = 1.253e-13); additionally, the correlation between the gene number per cluster and median V_A was negative (ρ = -0.277, p = 7.101e-05). The median V_D per cluster was positively correlated with the median transcript length (ρ = 0.201, p = 0.0042) and median gene exon number (ρ = 0.439, p = 7.752e-11). Accordingly, our data suggest that inter-allelic and inter-genic interactions both increase the amount of non-additive variance in gene expression.

Interestingly, there was a core set of 2,829 genes that always clustered together in the k-clustering analysis. Genes in this cluster had median V_D and V_A values of 0.501 and 0.141, respectively. GO enrichment analysis highlights 294 GO terms (significance threshold p = 0.0077; Supplementary Table S7). The top processes are related to developmental functions, such as "gravitropism" (p = 6.4e-28) and "trichome morphogenesis" (p = 8.9e-15).

Genes with high V_A in gene expression show higher levels of nucleotide variation

The random forest analysis indicated that none of the population genetic parameters of the original populations, such as F_{ST} (ρ = 0.0128, p = 0.164), had a significant impact on the level of non-additive variance (Figure 3a; Supplemental Table S3). In contrast, we observed that the amount of V_A was inversely correlated with genetic divergence between the two parental populations. The correlation was detected both at the individual gene level and at the level of gene clusters (Supplemental Table S4). The median V_A of each cluster was negatively correlated with the median value of relative population divergence F_{ST} (ρ = -0.201, p = 0.0041) and positively correlated with the absolute measure of population divergence d_{xy} (ρ = 0.320, p = 2.59e-06) among parental populations. The median V_A was also correlated with the median π values in both populations (SP: ρ = 0.34, p = 8.643e-07, PL: ρ = 0.33, p = 1.627e-07), and high V_A genes were among the genes that displayed the highest divergence at both synonymous and non synonymous sites (Supplementary Figure 3b-c). Taken together, our observations indicate that V_A evolves preferentially in transcripts displaying a higher mutation rate[42].

We used a generalized linear mixed model to test whether values of V_A, V_D, or V_G were predicted by F_{ST} and/or local signals of positive selection while controlling for gene identity. We also tested in the same model whether effects of selection signals differed among the three gene

sets. The variance distributions are all significantly different from each other ($p < 2.2e-16$), and the presence of a signature of a selective sweep in the vicinity of the expressed gene has a significant impact on the components of genetic variance of this gene ($p = 1.319e-4$). We further focused on the 156 SNPs displaying a significant association with the expression level of a gene. Because these SNPs did not co-locate with the expressed transcript, we can assume that they act as trans-regulatory variants. Overall, we find that when such a trans-regulatory SNP mapped to a genomic region carrying a signature of a selective sweep, the transcript it is associated with displayed significantly higher V_A and lower V_D than when this trans-regulator was located outside of a sweep area. Thus, recent selective sweeps seem to associate preferentially with higher additive variance in gene expression and to lower levels of non-additive variance. This observation accounts only for the 1% of transcripts for which a significant association could be detected, it thus does not contradict the observation reported above that genes with high V_A tend to show relaxed selective constraint. Clearly additive variance can sometimes be targeted by positive selection.

Discussion

Delineating the additive and non-additive components of genetic variance is crucial for understanding the adaptive potential of a trait, because only additive genetic variance can fuel a response to selection. Plants are particularly well suited for the dissection of genetic variance because both genetic crosses and environmental variance can be completely controlled; such control is necessary for a reliable estimation of the non-additive components of variance[13,43]. Our analysis of non-additive variance in 17,657 gene transcripts provides a unique opportunity to determine the factors that promote the emergence of variance in natural populations. Our data show that a substantial fraction of regulatory variation is not available to respond to selection despite being genetically determined. This work further documents three striking features of non-additive variance in gene expression: i) there is a strong relationship between mode of genetic inheritance of expression levels and the rate of amino acid evolution, ii) there are identifiable genomic features that favor non-additive variation in gene expression, and iii) non-additive variance is particularly elevated among genes related to the epigenetic regulation of gene expression.

First, we report a strong association between two independent sources of genetic variation: the composition of genetic variance in expression and the strength of selective constraint on amino-acid evolution. Gene regulation, which is central to the formation of phenotypes and variation in gene expression, is evidently subject to natural selection[21,44–47], and non-additive variance is expected to be higher in traits exposed to natural selection. A meta-analysis of non-additive variance in 338 traits in various organisms suggested that traits linked to fitness often have larger levels of non-additive variance in natural populations[16]. In maize, non-additive variance in 18 traits selected during domestication was also reported to be greater in the domesticated population than in its wild ancestor[43]. The scale of our study provides a compelling documentation of the relationship between the strength of selection and the amount of non-additive variance. Genes whose genetic variance at the transcript level is predominantly non-additive were significantly more constrained than genes showing the largest fraction of additive variance. The most strongly constrained genes, in turn, were found among the genes that were almost completely depleted of any genetic variance in transcript level, additive and non-additive variance alike. Although it has long been known that levels of genetic variation are low in genes with constrained expression[18,19,33–37,48], this study is, as far as we know, the first report that non-additive transcript variance increases when the strength of stabilizing selection increases but ultimately decreases in genes exposed to the strongest stabilizing selective forces. These differences are revealed by a skew in the frequency spectrum of non-synonymous mutations as well as in the number of amino-acid substitutions that ultimately fix in a population. The method used to estimate the DFE assumes that allelic effects are purely additive. If instead variants had recessive effects on fitness, the method would underestimate the magnitude of their effect and thus should lead to a pattern opposite to the one we observe, where variants in high V_A genes experienced weaker selection than variants in high V_D genes. Our results are therefore robust to hypothesized fitness effects of alleles. Although non-additive genetic variance emerges as a result of intra- and inter-locus interactions, no such direct relationship exists between allelic effects and the amount of non-additive variance[3]. This is because non-additive variance is a property of the population resulting from the frequency at which genetic variants segregate as well as the molecular interactions between them[13,49]. It thus would not make much sense to model the DFE assuming distinct values of the allelic dominance h for the different groups of genes. We note however, that our data reveals an important additional feature: the rate of

synonymous divergence and polymorphism is significantly lower in genes with high non-additive variance compared to all other genes, suggesting that background selection or the mutation rate may also play a role on the composition of genetic variance for expression.

Second, our data indicate that non-additive variance is promoted by transcript length, suggesting that the mutational target size of the expressed locus favors the emergence of such variance. Our study of transcript-level variance further indicates that non-additive variance increases with the size of the network of co-regulated genes, which confirms the observation that proteins that are more central tend to be more constrained at the amino-acid level[15,50–52]. Nevertheless, this work also shows that what promotes non-additive phenotypic variance is distinct from the molecular interaction between alleles. Indeed, additive molecular effects on fitness (e.g. co-dominant alleles) were inferred, in the same species, to be more frequent in highly connected genes, a result opposite to our findings[53].

Finally, we report that non-additive variance is particularly predominant among genes influencing the epigenetic remodeling of chromatin in differentiating cells. Interestingly, we also associated non-additive variance in expression with the presence of sites bound by a regulator of cell differentiation associated with stem thickening during growth[54]. Although epigenetic variation has been often hypothesized to facilitate adaptation[55], our results indicate that given the genetic architecture of epigenetic regulation, it is comparatively less likely to produce the additive variance needed to fuel adaptation in *A. lyrata*. Differentiating between additive and non-additive sources of variance appears to be crucial to understand the role of gene expression variation for past and future adaptation.

Acknowledgments

We would like to thank Kirsten Bell for help with the laboratory work, the Cologne Genome Center for the transcriptome sequences and Emily Wheeler, Harwich, MA, for editorial help. Furthermore, we would like to thank F. He, S. Sunyaev, O. Savolainen, T. Hämälä, and M. Stetter for feedback, and F. Hospital for inspiring discussions ahead of this project. This research was funded by ERC Grant No. 648617 Adaptoscope.

Data availability

The whole-genome and transcriptome sequences are available via ENA. The whole-genome sequences of the parental lines are under the number PRJEB33206, and the transcriptome sequences are under the number PRJEB49153. The gene counts per individual and the sql database compiling descriptors of genomic variation will be deposited in dryad upon acceptance of the manuscript. All the original scripts are in git repository https://github.com/mtakou/Quantitative-analysis-of-gene-expression.

References

1. Savolainen, O., Lascoux, M. & Merilä, J. Ecological genomics of local adaptation. *Nat Rev Genet* **14**, 807–820 (2013).
2. Fisher, R. A. The correlation between relatives on the supposition of mendelian inheritance. *Trans. R. Soc. Edinb* **53**, 399–433 (1918).
3. Huang, W. & Mackay, T. F. C. The Genetic Architecture of Quantitative Traits Cannot Be Inferred from Variance Component Analysis. *PLOS Genetics* **12**, e1006421 (2016).
4. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).
5. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell* **169**, 1177–1186 (2017).
6. Zhu, Z. *et al.* Dominance genetic variation contributes little to the missing heritability for human complex traits. *Am J Hum Genet* **96**, 377–385 (2015).
7. Hill, W. G., Goddard, M. E. & Visscher, P. M. Data and Theory Point to Mainly Additive Genetic Variance for Complex Traits. *PLoS Genet* **4**, e1000008 (2008).
8. Crow, J. F. On epistasis: why it is unimportant in polygenic directional selection. *Philosophical Transactions of the Royal Society B: Biological Sciences* **365**, 1241–1244 (2010).
9. Visscher, P. M., Hill, W. G. & Wray, N. R. Heritability in the genomics era — concepts and misconceptions. *Nature Reviews Genetics* **9**, 255–266 (2008).

380 10. Mäki-Tanila, A. & Hill, W. G. Influence of Gene Interaction on Complex Trait Variation with
381 Multilocus Models. *Genetics* **198**, 355–367 (2014).
382 11. Willis, J. H. & Orr, H. A. Increased Heritable Variation Following Population Bottlenecks: The Role of
383 Dominance. *Evolution* **47**, 949–957 (1993).
384 12. Wang, J., Caballero, A., Keightley, P. D. & Hill, W. G. Bottleneck effect on genetic variance. A
385 theoretical investigation of the role of dominance. *Genetics* **150**, 435–447 (1998).
386 13. Lynch, M. & Walsh, B. *Genetics and Analysis of Quantitative Traits*. (Sinauer Associates, Sunderland,
387 Massachusetts, 1998).
388 14. Mousseau, T. A. & Roff, D. A. Natural selection and the heritability of fitness components. *Heredity*
389 **59**, 181–197 (1987).
390 15. Orr, H. A. Fitness and its role in evolutionary genetics. *Nat Rev Genet* **10**, 531–539 (2009).
391 16. Crnokrak, P. & Roff, D. A. Dominance variance: associations with selection and fitness. *Heredity* **75**,
392 530–540 (1995).
393 17. Waldmann, P. Additive and non-additive genetic architecture of two different-sized populations of
394 Scabiosa canescens. *Heredity* **86**, 648–657 (2001).
395 18. Ayroles, J. F. *et al.* Systems genetics of complex traits in Drosophila melanogaster. *Nat Genet* **41**,
396 299–307 (2009).
397 19. Romero, I. G., Ruvinsky, I. & Gilad, Y. Comparative studies of gene expression and the evolution of
398 gene regulation. *Nat Rev Genet* **13**, 505–516 (2012).
399 20. Oleksiak, M. F., Churchill, G. A. & Crawford, D. L. Variation in gene expression within and among
400 natural populations. *Nature Genetics* **32**, 261–266 (2002).
401 21. Groen, S. C. *et al.* The strength and pattern of natural selection on gene expression in rice. *Nature*
402 1–5 (2020) doi:10.1038/s41586-020-1997-2.

403 22. Leder, E. H. *et al.* The Evolution and Adaptive Potential of Transcriptional Variation in Sticklebacks—
404 Signatures of Selection and Widespread Heritability. *Mol Biol Evol* **32**, 674–689 (2015).
405 23. Mattila, T. M., Tyrmi, J., Pyhäjärvi, T. & Savolainen, O. Genome-Wide Analysis of Colonization
406 History and Concomitant Selection in Arabidopsis lyrata. *Mol Biol Evol* **34**, 2665–2677 (2017).
407 24. Hämälä, T., Mattila, T. M. & Savolainen, O. Local adaptation and ecological differentiation under
408 selection, migration, and drift in Arabidopsis lyrata. *Evolution* **72**, 1373–1386 (2018).
409 25. Takou, M. *et al.* Maintenance of Adaptive Dynamics and No Detectable Load in a Range-Edge
410 Outcrossing Plant Population. *Molecular Biology and Evolution* **38**, 1820–1836 (2021).
411 26. Reif, J. C., Gumpert, F.-M., Fischer, S. & Melchinger, A. E. Impact of Interpopulation Divergence on
412 Additive and Dominance Variance in Hybrid Populations. *Genetics* **176**, 1931–1934 (2007).
413 27. Bullard, J. H., Mostovoy, Y., Dudoit, S. & Brem, R. B. Polygenic and directional regulatory evolution
414 across pathways in Saccharomyces. *PNAS* **107**, 5058–5063 (2010).
415 28. Lafon-Placette, C. & Köhler, C. Epigenetic mechanisms of postzygotic reproductive isolation in
416 plants. *Current Opinion in Plant Biology* **23**, 39–44 (2015).
417 29. Todesco, M. *et al.* Hybridization and extinction. *Evolutionary Applications* **9**, 892–908 (2016).
418 30. Alcázar, R. *et al.* Natural variation at Strubbelig Receptor Kinase 3 drives immune-triggered
419 incompatibilities between Arabidopsis thaliana accessions. *Nat Genet* **42**, 1135–1139 (2010).
420 31. Bomblies, K. *et al.* Local-Scale Patterns of Genetic Variability, Outcrossing, and Spatial Structure in
421 Natural Stands of Arabidopsis thaliana. *PLOS Genetics* **6**, e1000890 (2010).
422 32. He, F. *et al.* Cis-regulatory evolution spotlights species differences in the adaptive potential of gene
423 expression plasticity. *Nat Commun* **12**, 3376 (2021).
424 33. Fay, J. C. & Wittkopp, P. J. Evaluating the role of natural selection in the evolution of gene
425 regulation. *Heredity* **100**, 191–199 (2008).

426 34. Steige, K. A., Reimegård, J., Koenig, D., Scofield, D. G. & Slotte, T. Cis-Regulatory Changes Associated
427 with a Recent Mating System Shift and Floral Adaptation in Capsella. *Mol Biol Evol* **32**, 2501–2514
428 (2015).
429 35. Hämälä, T. *et al.* Gene Expression Modularity Reveals Footprints of Polygenic Adaptation in
430 Theobroma cacao. *Mol Biol Evol* **37**, 110–123 (2020).
431 36. Whitehead, A. & Crawford, D. L. Neutral and adaptive variation in gene expression. *PNAS* **103**,
432 5425–5430 (2006).
433 37. Hodgins-Davis, A., Rice, D. P. & Townsend, J. P. Gene Expression Evolves under a House-of-Cards
434 Model of Stabilizing Selection. *Mol Biol Evol* **32**, 2130–2140 (2015).
435 38. Metzger, B. P. H., Wittkopp, P. J. & Coolon, J. D. Evolutionary Dynamics of Regulatory Changes
436 Underlying Gene Expression Divergence among Saccharomyces Species. *Genome Biol Evol* **9**, 843–
437 854 (2017).
438 39. Khaitovich, P., Enard, W., Lachmann, M. & Pääbo, S. Evolution of primate gene expression. *Nat Rev*
439 *Genet* **7**, 693–702 (2006).
440 40. Schrider, D. R. Background Selection Does Not Mimic the Patterns of Genetic Diversity Produced by
441 Selective Sweeps. *Genetics* **216**, 499–519 (2020).
442 41. Miyashima, S. *et al.* Mobile PEAR transcription factors integrate positional cues to prime cambial
443 growth. *Nature* **565**, 490–494 (2019).
444 42. Cruickshank, T. E. & Hahn, M. W. Reanalysis suggests that genomic islands of speciation are due to
445 reduced diversity, not reduced gene flow. *Mol Ecol* **23**, 3133–3157 (2014).
446 43. Yang, C. J. *et al.* The genetic architecture of teosinte catalyzed and constrained maize domestication.
447 *Proc Natl Acad Sci USA* **116**, 5643–5652 (2019).
448 44. Kremling, K. A. G. *et al.* Dysregulation of expression correlates with rare-allele burden and fitness
449 loss in maize. *Nature* **555**, 520–523 (2018).

450 45. Studer, A., Zhao, Q., Ross-Ibarra, J. & Doebley, J. Identification of a functional transposon insertion
451 in the maize domestication gene tb1. *Nat Genet* **43**, 1160–1163 (2011).
452 46. Clark, R. M., Wagler, T. N., Quijada, P. & Doebley, J. A distant upstream enhancer at the maize
453 domestication gene tb1 has pleiotropic effects on plant and inflorescent architecture. *Nat Genet* **38**,
454 594–597 (2006).
455 47. Jacob, F. & Monod, J. Genetic regulatory mechanisms in the synthesis of proteins. *Journal of*
456 *Molecular Biology* **3**, 318–356 (1961).
457 48. Bedford, T. & Hartl, D. L. Optimization of gene expression by natural selection. *PNAS* **106**, 1133–
458 1138 (2009).
459 49. Falconer, D. S. & Mackay, T. F. C. *Introduction to Quantitative Genetics*. (Prentice Hall, 1996).
460 50. Hahn, M. W. & Kern, A. D. Comparative Genomics of Centrality and Essentiality in Three Eukaryotic
461 Protein-Interaction Networks. *Mol Biol Evol* **22**, 803–806 (2005).
462 51. Josephs, E. B., Wright, S. I., Stinchcombe, J. R. & Schoen, D. J. The Relationship between Selection,
463 Network Connectivity, and Regulatory Variation within a Population of Capsella grandiflora. *Genome*
464 *Biol Evol* **9**, 1099–1109 (2017).
465 52. Masalia, R. R., Bewick, A. J. & Burke, J. M. Connectivity in gene coexpression networks negatively
466 correlates with rates of molecular evolution in flowering plants. *PLOS ONE* **12**, e0182289 (2017).
467 53. Huber, C. D., Durvasula, A., Hancock, A. M. & Lohmueller, K. E. Gene expression drives the evolution
468 of dominance. *Nature Communications* **9**, 2750 (2018).
469 54. Lister, R. *et al.* A Transcriptomic and Proteomic Characterization of the Arabidopsis Mitochondrial
470 Protein Import Apparatus and Its Response to Mitochondrial Dysfunction. *Plant Physiol* **134**, 777–
471 789 (2004).
472 55. Schmid, M. W. *et al.* Contribution of epigenetic variation to adaptation in Arabidopsis. *Nat Commun*
473 **9**, 4446 (2018).

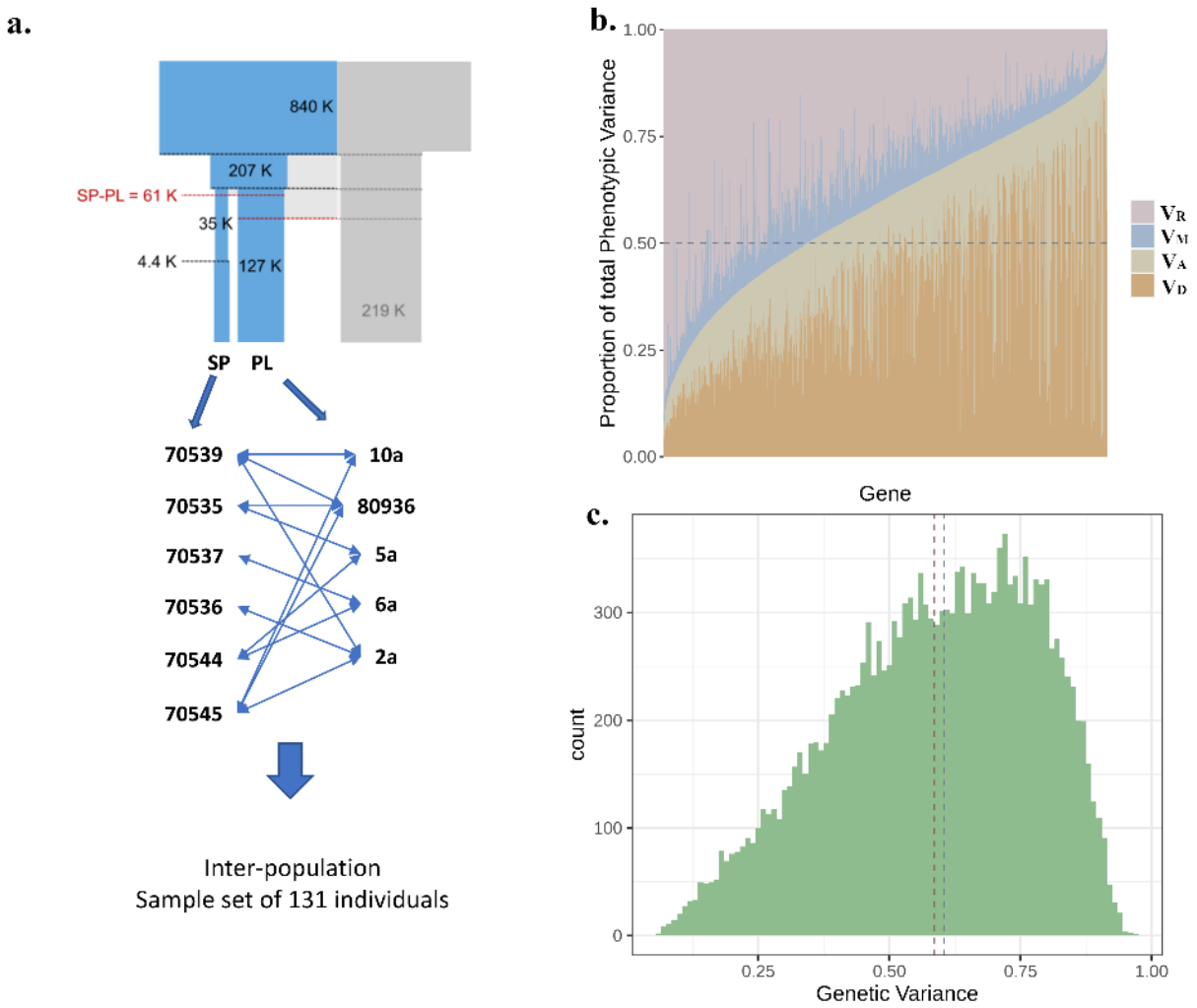

Figure 1: Phenotypic variance of gene expression is mostly comprised of non-additive genetic variance. **a.** We used 2 populations of *A. lyrata*, performed a set of inter-population crosses, and generated a set of 131 individuals. The two populations separated from each other approximately 61 kyears ago, with the SP populations being more isolated and more severely bottlenecked than the PL population (redrawn from [25]). In total, 5 individuals sampled in Plech (PL) were reciprocally crossed with 6 individuals from Spiterstulen (SP). The crosses were done in such a way that plants were reciprocally crossed with 2 individuals from the other population in order to produce full- and half-sib families. **b.** The total phenotypic variance of each one of the 17,656 genes included in the analysis were partitioned into non-additive, additive, maternal, and residual components of variance. The genes were ordered by increasing values of the fraction of total phenotypic variance in transcript level that was attributed to genetic variance (V_G). The dashed line marks the 0.5 point where half of the variance is explained. **c.** The distribution of total genetic variance values for each transcript (V_G), as a fraction of the total observed variance in expression level. The dashed blue line shows the median value (0.604) of the distribution; the dashed red line indicates the mean value (0.585).

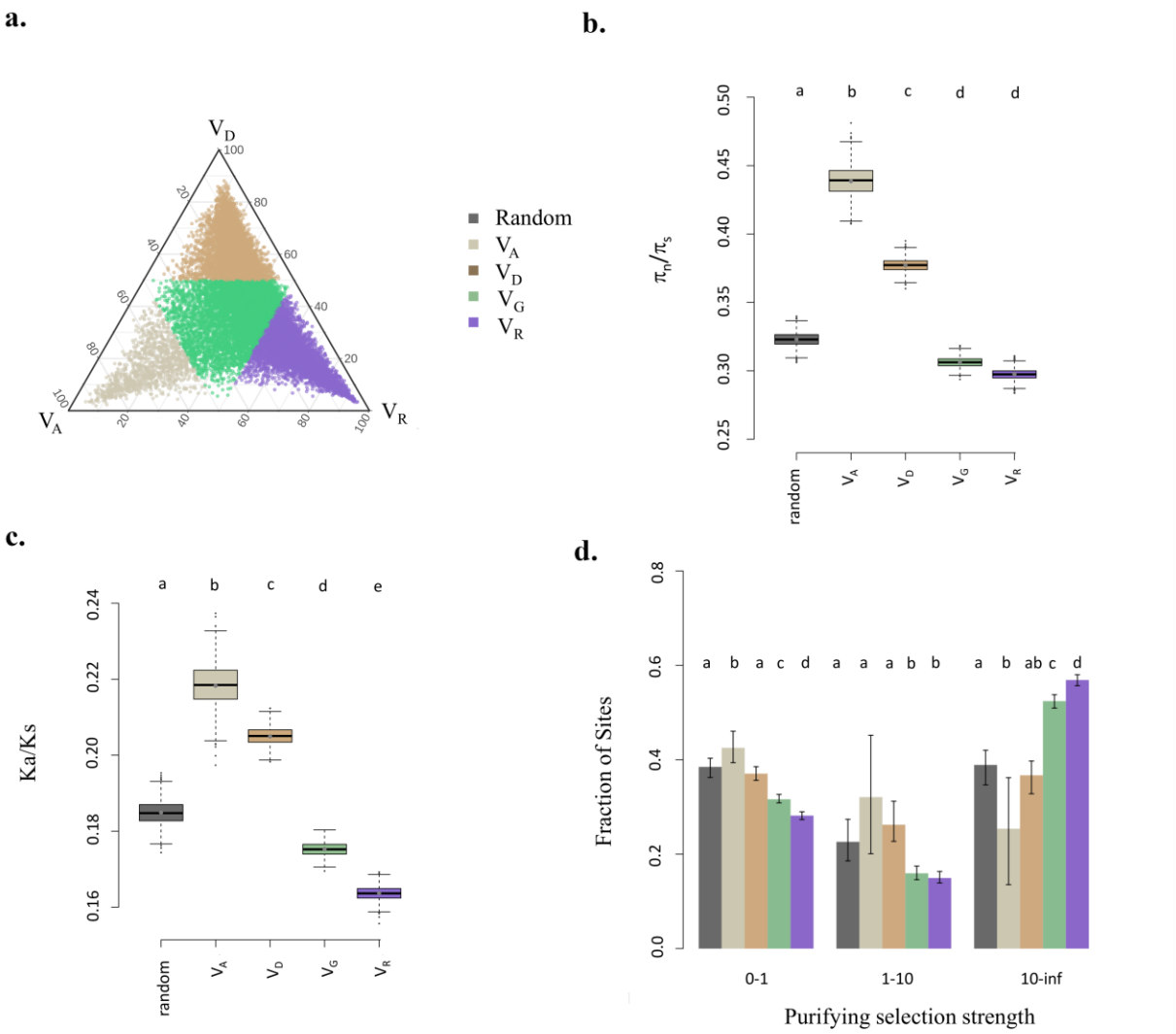

Figure 2: Constraints on the amino-acid level depend on the predominant component of genetic variance. **a.** We grouped the genes based on the partitioning of total phenotypic variance in transcript level into non-additive, additive, or residual variance. The threshold for each case is 50% of the total phenotypic variance. **b.** We computed the mean and variance of average pi/ps values for each group based on population genomics datasets for PL. **c.** The Ka/Ks ratio for the different gene groups. The confidence intervals and *p* values for b through d were estimated based on 100 bootstrap iterations. **d.** Inferred distribution of fitness effects of new mutations in the four gene groups. Gray: control group of 3500 random genes; beige: high V_A genes; orange: high V_D genes; purple: high V_R genes; green: intermediate levels of V_G. In all plots, the letters indicate whether the values differ significantly from each other.

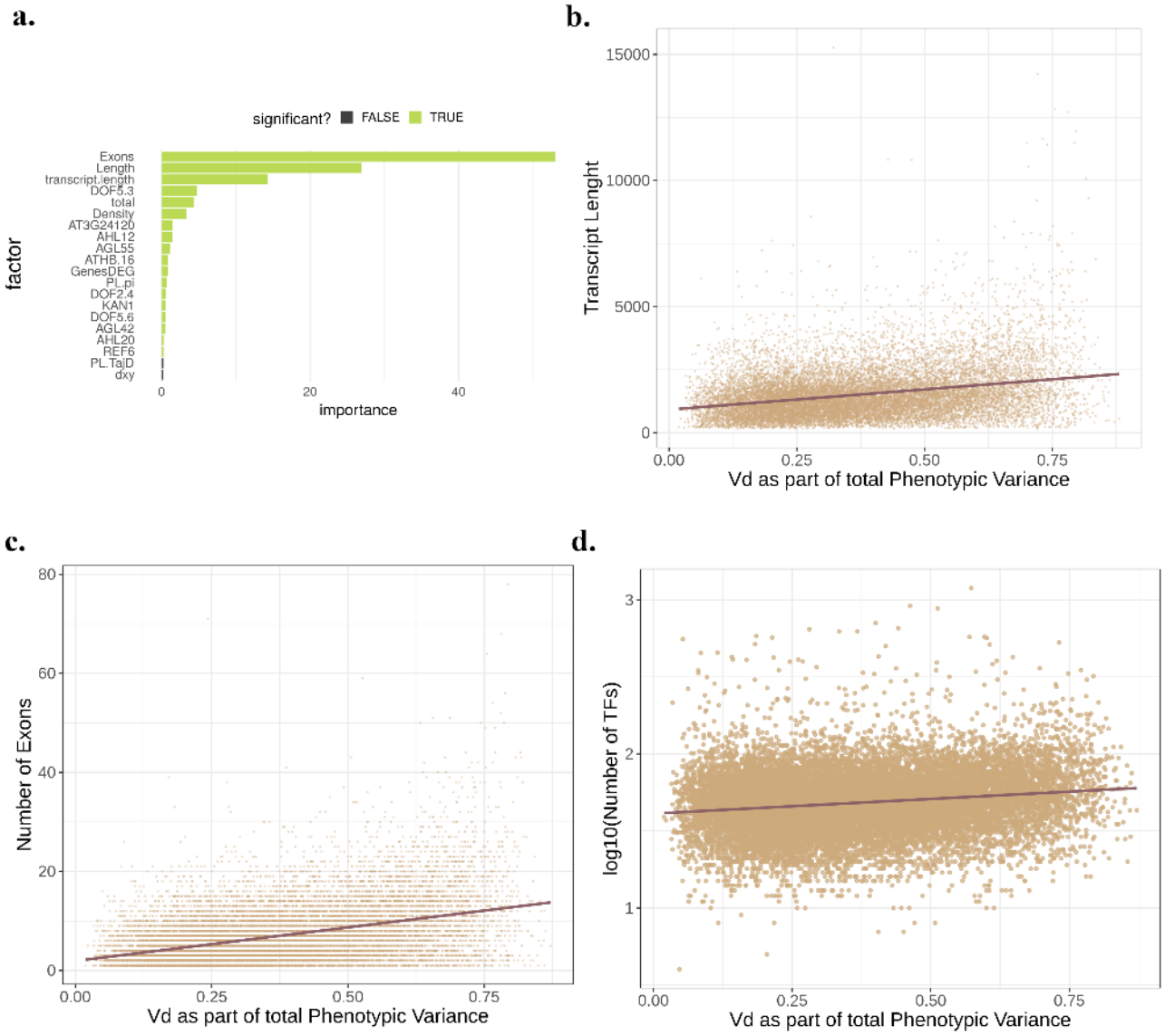

Figure 3: Gene architecture controls the level of non-additive variance. **a.** The significance and importance of tested factors in the random forest analysis that explains the level of non-additive variance each gene has. The twenty factors that explain the most variance are shown. The predictive power of the model is $R^2 = 0.18$. **b.** The non-additive variance correlates significantly with the transcript length of each gene (p=1.22e-266, ρ=0.258). **c.** The fraction of non-additive variance correlates significantly with the number of exons that a gene has (p < 2.2e-16, ρ=0.385). **d.** The non-additive variance correlates significantly with the total number of transcription-factor binding sites detected in the upstream vicinity of the gene (p < 2.2e-16, ρ=0.162).