Supplementary Information

S1 Canonical Correlation Analysis

Given two sets of random variables $(X_1, X_2, \dots, X_p)$ and $(Y_1, Y_2, \dots, Y_q)$, CCA finds linear coefficients $a \in \mathbb{R}^p$ and $b \in \mathbb{R}^q$ that maximize the correlation

$$\rho(a,b) = \mathrm{corr}\left(\sum_{i=1}^{p} a_i X_i, \sum_{j=1}^{q} b_j Y_j\right)$$

It can be shown [1] that the optimal vector $\mathbf{a}$ is an eigenvector of the matrix $\Sigma_{XX}^{-1}\Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX}$, where Σ_{XX}, Σ_{XY} and Σ_{YY} are the covariance matrices among the X and Y variables. In the special case where $q = 1$ (one SNP), Σ_{YY} is a number and Σ_{XY} a column vector, and this matrix takes the form $\Sigma_{XX}^{-1}\mathbf{v}\mathbf{v}^T$, where $\mathbf{v} = \Sigma_{YY}^{-1}\Sigma_{XY}$. The (only) eigenvector of such a matrix is $\mathbf{a} = \Sigma_{XX}^{-1}\mathbf{v}$.

To estimate the coefficients $\mathbf{a}$ from data, assume that we have standardized data $\mathbf{X} \in \mathbb{R}^{n\times p}$ and $\mathbf{y} \in \mathbb{R}^n$, such that

$$\sum_{k=1}^{n} x_{ik} = \sum_{k=1}^{n} y_k = 0 \qquad\qquad \frac{1}{n-1}\sum_{k=1}^{n} x_{ik}^2 = \frac{1}{n-1}\sum_{k=1}^{n} y_k^2 = 1.$$

Then the estimates for the covariances are

$$\hat{\Sigma}_{XX} = \frac{\mathbf{X}^T\mathbf{X}}{n-1} \qquad \hat{\Sigma}_{XY} = \frac{\mathbf{X}^T\mathbf{y}}{n-1} \qquad \hat{\Sigma}_{YY} = 1,$$

and hence

$$\hat{\mathbf{a}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}.$$

S2 Multi-trait GWAS approaches

As in other multi-trait GWAS methods, we consider one genetic variant at a time, and represent it by a random variable Y. We consider p traits represented by random variables $X_1, X_2, \dots, X_p$ taking real values. We define a “forward” multi-trait association model probabilistically through a conditional distribution $p(X_1, \dots, X_p \mid Y)$, which corresponds to the natural direction where variation in Y causes variation in the X_i. Using Bayes’ formula, we can write the same model in the reverse causal direction using Y as the dependent variable:

$$P(Y \mid X_1, \dots, X_p) = P(X_1, \dots, X_p \mid Y)\frac{P(Y)}{P(X_1, \dots, X_p)} \tag{1}$$

[1] See for instance
Hardoon DR, Szedmak S and Shawe-Taylor J. Canonical correlation analysis: An overview with application to learning methods *Neural computation* **16**:2639–2664 (2003).

where $P(Y)$ and $P(X_1, \dots, X_p)$ are prior distributions. Conversely, a forward model $P(X_1, \dots, X_p \mid Y)$ can be obtained from a reverse model $P(Y \mid X_1, \dots, X_p)$ using the same formula.

We have data in the form of independent random samples from the joint distribution $P(Y, X_1, \dots, X_p)$ in n individuals, represented by a genotype vector $\mathbf{y} \in \mathbb{R}^n$ and trait vectors $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_p \in \mathbb{R}^n$, which we gather in a matrix $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_p) \in \mathbb{R}^{n \times p}$. The log-likelihood of observing the data is the log-probability density

$$\begin{aligned}\mathcal{L} = \log p(\mathbf{y}, \mathbf{X}) &= \log \prod_{j=1}^{n} p(y_j, x_{j1}, \dots, x_{jp}) \\ &= \sum_{j=1}^{n} \log p(y_j, x_{j1}, \dots, x_{jp}),\end{aligned}$$

which can be expressed in terms of the forward or reverse conditional probabilities depending on the type of model being fit. We now review how existing as well as newly proposed, and low-dimensional as well as high-dimensional multi-trait GWAS methods fit within this framework.

S2.1 Univariate tests

The simplest method for multi-trait GWAS in the high-dimensional setting consists of testing each trait for association with the genetic variant independently. In this case we fit, by maximum-likelihood, a model $p(x_i \mid y)$ for each trait X_i independently using a linear model

$$p(x_i \mid y) = \mathcal{N}(\mu_y, \sigma^2)$$

a normal distribution with mean μ_y dependent on the genotype value y. This corresponds to the multi-trait model

$$p(x_1, \dots, x_p \mid y) = \prod_{i=1}^{p} p(x_i \mid y)$$

Using Bayes' rule eq. (1), we obtain

$$\begin{aligned}P(y \mid x_1, \dots, x_p) &= p(x_1, \dots, x_p \mid y) \frac{P(y)}{p(x_1, \dots, x_p)} \\ &\propto P(y) \prod_{i=1}^{p} p(x_i \mid y)\end{aligned}$$

where $P(y)$ is the prior probability (background frequency) of observing genotype class y. This is the formula for a *naive Bayes classifier* of the genotype y given features x_i. In the univariate approach, statistical tests are carried out to determine whether a genotype-dependent model $p(x_i \mid y)$ is more likely or not than a model where the trait is independent of the genotype. This is equivalent to doing a feature selection to determine which traits to include in the naive Bayes classifier.

S2.2 Canonical correlation analysis

MV-PLINK [2] is a multivariate method based on Canonical Correlation Analysis (CCA). Given two sets of random variables $(X_1, X_2, \dots, X_p)$ and
$(Y_1, Y_2, \dots, Y_q)$, CCA finds linear coefficients $\mathbf{a} \in \mathbb{R}^p$ and $\mathbf{b} \in \mathbb{R}^q$ that maximize the correlation

$$\rho(\mathbf{a}, \mathbf{b}) = \text{corr}\left(\sum_{i=1}^{p} a_i X_i, \sum_{j=1}^{q} b_j Y_j\right)$$

It can be shown (see SI Section S1) that if $q = 1$, then the maximizing coefficients $\hat{\mathbf{a}}$ are given by $\hat{\mathbf{a}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$, where $\mathbf{X}$ and $\mathbf{y}$ are the data sampled from the joint distribution $P(Y, X_1, X_2, \dots, X_p)$. These are the same coefficients that would be obtained from a *linear regression* model where Y is modelled as a linear function of the predictors $(X_1, X_2, \dots, X_p)$, or from the maximum-likelihood solution of a reverse probabilistic model

$$p(y \mid x_1, \dots, x_p) = \mathcal{N}\left(\sum_{i=1}^{p} a_i x_i, \sigma^2\right). \tag{2}$$

S2.3 Reverse logistic regression

MultiPhen [3] is a method that is described directly in terms of a model to predict genotypes from multiple traits, using proportional odds *logistic regression*, that is, instead of fitting the genotype class probabilities $P(y = m \mid x_1, \dots, x_p)$, for $m = 0, 1, 2$ (for biallelic data), the method fits

$$P(y \leq m \mid x_1, \dots, x_p) = \frac{1}{1 + e^{-\alpha_m - \sum_{i=1}^{p} \beta_i x_i}}$$

Then a likelihood ratio test is used to determine if this model fits the data better than a model where $\beta_1 = \dots = \beta_p = 0$, thus carrying out a single test for each genetic variant, testing whether the variant is associated with *any* of the traits using the logistic regression model.

S2.4 L2-Regularized reverse regression

Expressing CCA for multi-trait GWAS as a linear regression of the variant genotype on the trait values [eq. (2)] immediately leads to a generalization to the high-dimensional setting in the form of regularizing the regression coefficients, that is, augmenting eq. (2) with a prior distribution $p(a_i) = \mathcal{N}(0, \alpha)$, $i = 1, \dots, p$. Finding the maximum-likelihood values of the regression coefficients is equivalent to L_2-regularized or ridge regression. This is the approach followed by [**?**], who combined it with a likelihood ratio test to

[2] Ferreira MA, Purcell SM, multivariate test of association. *Bioinformatics* **25**(1):132–133 (2009).

[3] O'Reilly PF, et al., Multi- Phen: joint model of multiple phenotypes can increase discovery in GWAS. *PloS one* **7**(5):e34861 (2012).

determine whether the fitted model is more likely than a model where the genotype is independent of the traits ($a_i = 0$ for all i) and obtain a single association p-value for each variant.

S2.5 Reverse genotype prediction using machine learning methods

From the above, we conclude that existing multi-trait GWAS methods can be described as reverse genotype prediction methods. From this perspective, L_2-regularized linear regression is but one of many established machine learning methods that could be used to predict an outcome variable Y from a high number of predictors or features X_i, $i = 1, \ldots, p$. Hence we propose to consider a wider range of machine learning methods such as random forest regression (RFR) and support vector regression (SVR) [4]. Our overall hypothesis is that genetic variants whose genotypes can be predicted with higher accuracy are more likely to affect some or all of the traits under consideration than variants whose genotypes cannot be predicted well, and that feature weights in the fitted models measure the strength of biological association between a variant and a trait.

S3 Supplementary Figures

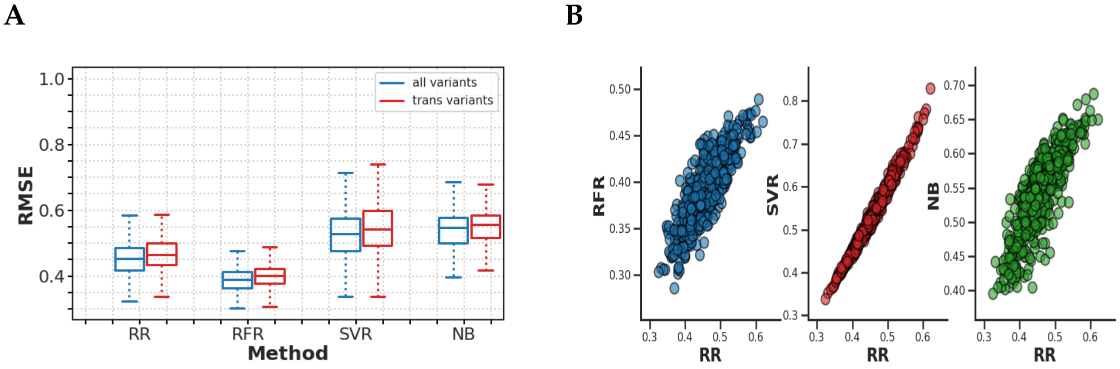

Figure S1: RMSE values for genotype prediction on DREAM5 simulated data. **A.** Boxplots show the distribution of the RMSE values for all variants (blue) and for trans-acting-only variants (red) for random forest regression (RFR), support vector regression (SVR), ridge regression (RR), and naive Bayes (NB). **B.** Scatter plots show RMSE values of RFR, SVR, and NB vs RR for all variants. The data shown are for **DREAM Network 2**.

[4] Friedman J, et al., The elements of statistical learning *Springer series in statistics*, **vol. 1**, NewYork(2001).

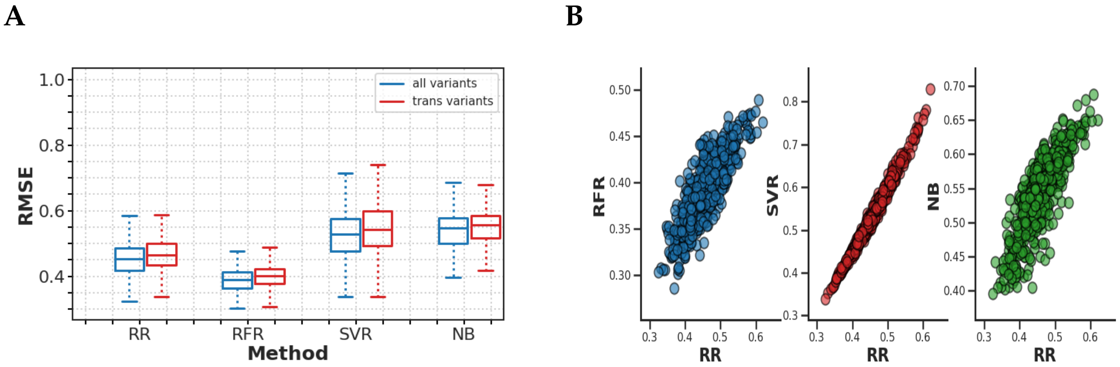

Figure S2: RMSE values for genotype prediction on DREAM5 simulated data. **A.** Box-plots show the distribution of the RMSE values for all variants (blue) and for trans-acting-only variants (red) for random forest regression (RFR), support vector regression (SVR), ridge regression (RR), and naive Bayes (NB). **B.** Scatter plots show RMSE values of RFR, SVR, and NB vs RR for all variants. The data shown are for **DREAM Network 3**.

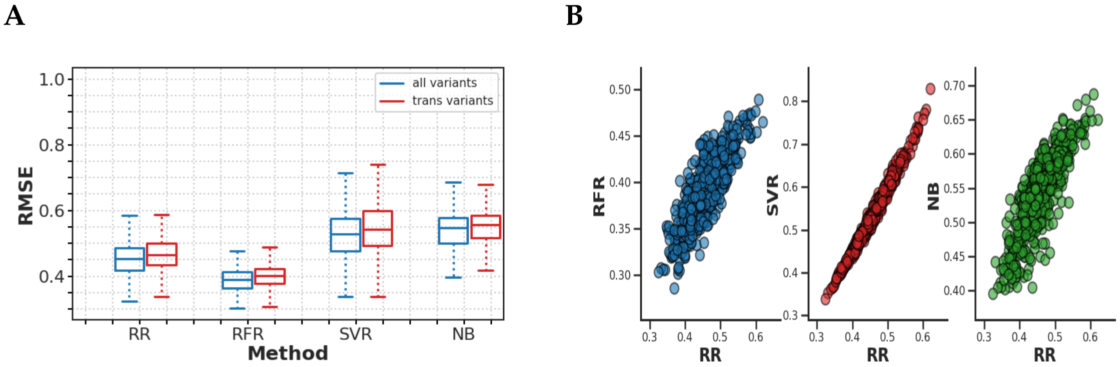

Figure S3: RMSE values for genotype prediction on DREAM5 simulated data. **A.** Box-plots show the distribution of the RMSE values for all variants (blue) and for trans-acting-only variants (red) for random forest regression (RFR), support vector regression (SVR), ridge regression (RR), and naive Bayes (NB). **B.** Scatter plots show RMSE values of RFR, SVR, and NB vs RR for all variants. The data shown are for **DREAM Network 4**.

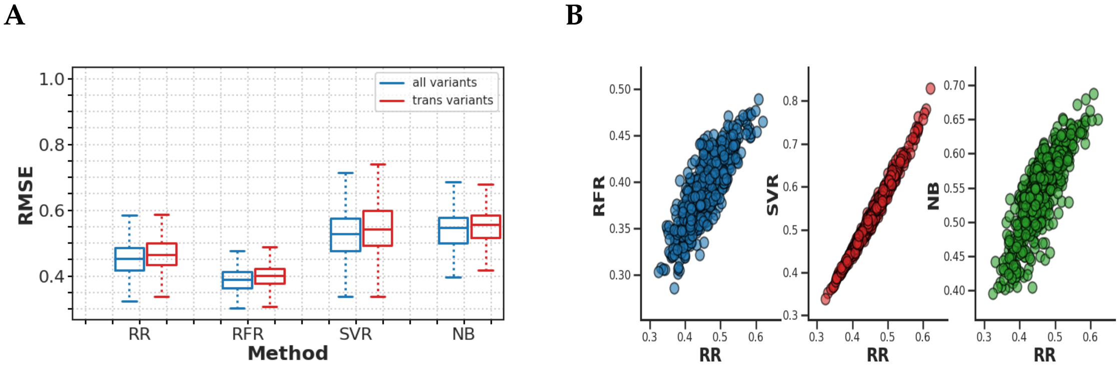

Figure S4: RMSE values for genotype prediction on DREAM5 simulated data. **A.** Box-plots show the distribution of the RMSE values for all variants (blue) and for trans-acting-only variants (red) for random forest regression (RFR), support vector regression (SVR), ridge regression (RR), and naive Bayes (NB). **B.** Scatter plots show RMSE values of RFR, SVR, and NB vs RR for all variants. The data shown are for **DREAM Network 5**.

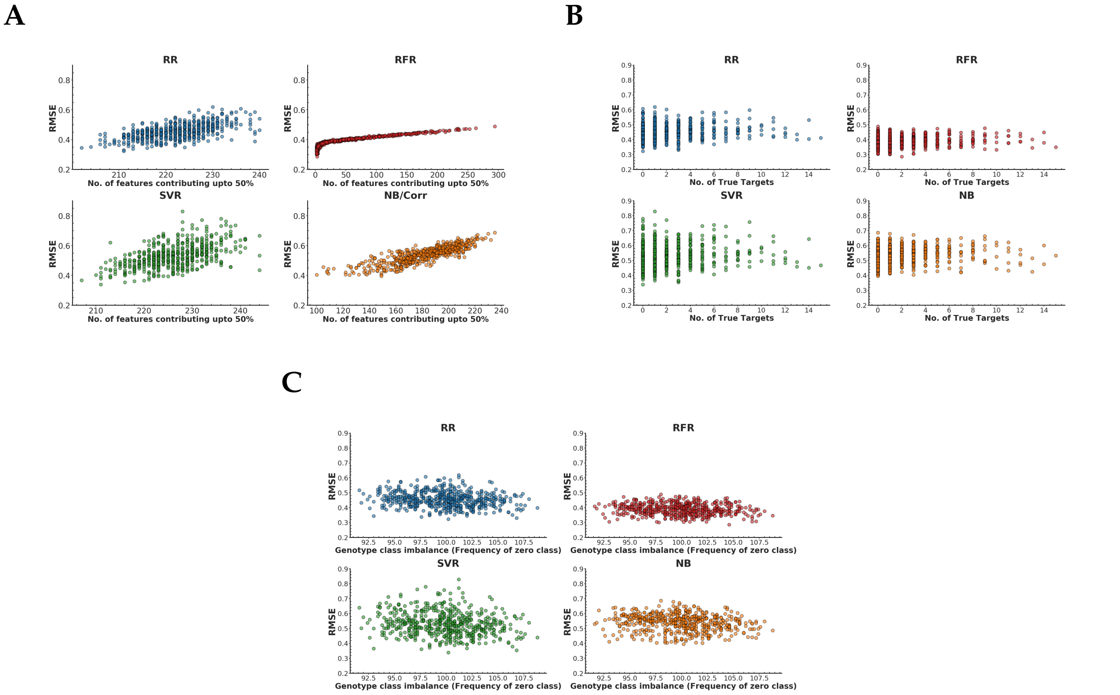

Figure S5: Scatter plots of genotype RMSE values on DREAM5 simulated data against the number of selected model features (**A**), the number of true trans-eQTL targets in the ground-truth network (**B**), and the genotype class balance (frequency of the zero class) (**C**), for random forest regression (RFR), support vector regression (SVR), ridge regression (RR), and naive Bayes (NB). The data shown are for **DREAM Network 2**.

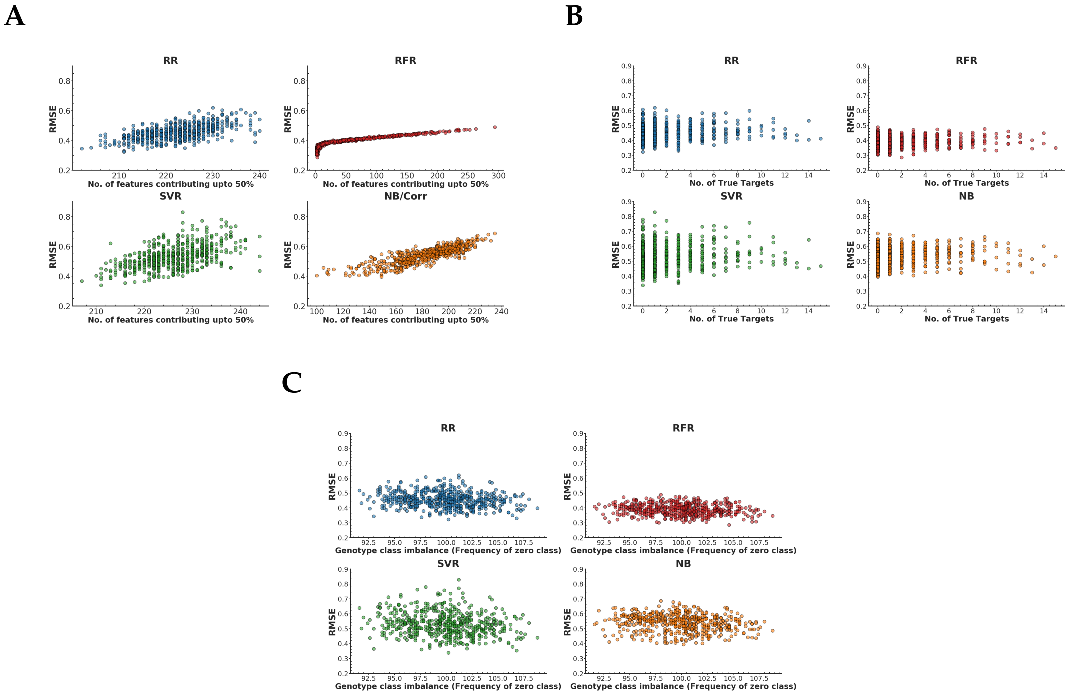

Figure S6: Scatter plots of genotype RMSE values on DREAM5 simulated data against the number of selected model features (**A**), the number of true trans-eQTL targets in the ground-truth network (**B**), and the genotype class balance (frequency of the zero class) (**C**), for random forest regression (RFR), support vector regression (SVR), ridge regression (RR), and naive Bayes (NB). The data shown are for **DREAM Network 3**.

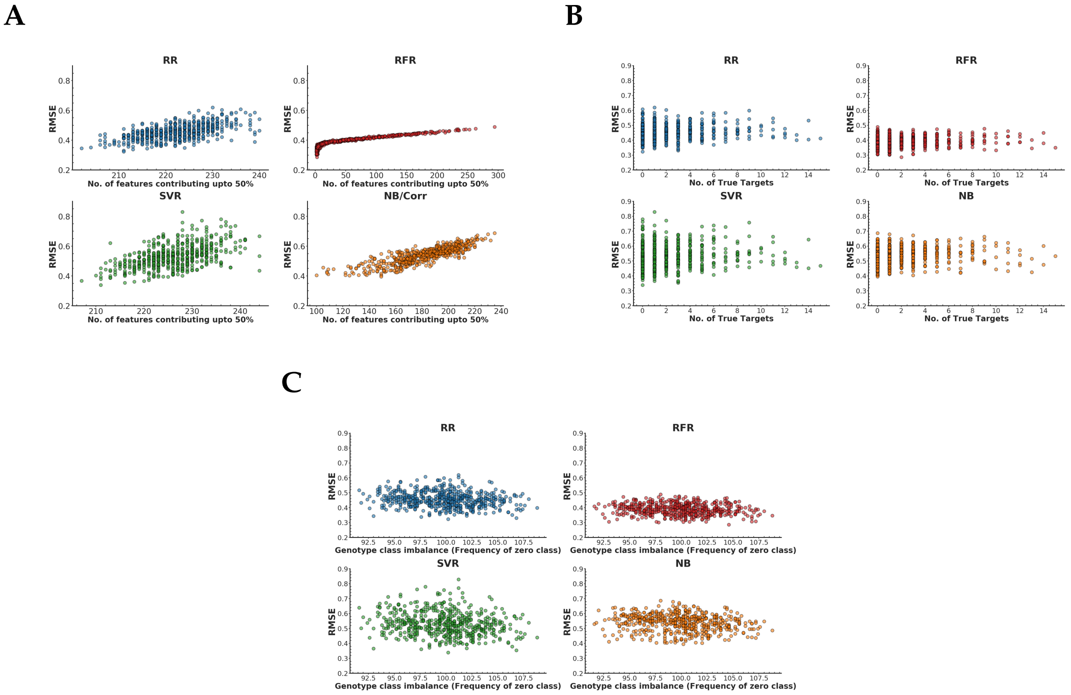

Figure S7: Scatter plots of genotype RMSE values on DREAM5 simulated data against the number of selected model features (**A**), the number of true trans-eQTL targets in the ground-truth network (**B**), and the genotype class balance (frequency of the zero class) (**C**), for random forest regression (RFR), support vector regression (SVR), ridge regression (RR), and naive Bayes (NB). The data shown are for **DREAM Network 4**.

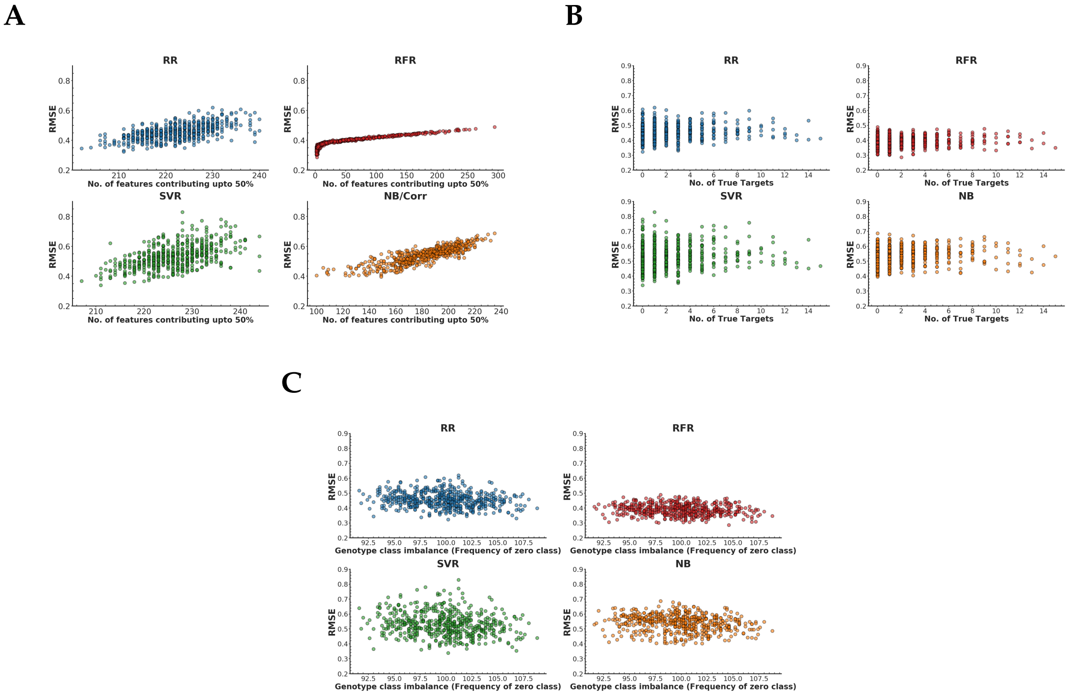

Figure S8: Scatter plots of genotype RMSE values on DREAM5 simulated data against the number of selected model features (**A**), the number of true trans-eQTL targets in the ground-truth network (**B**), and the genotype class balance (frequency of the zero class) (**C**), for random forest regression (RFR), support vector regression (SVR), ridge regression (RR), and naive Bayes (NB). The data shown are for **DREAM Network 5**.

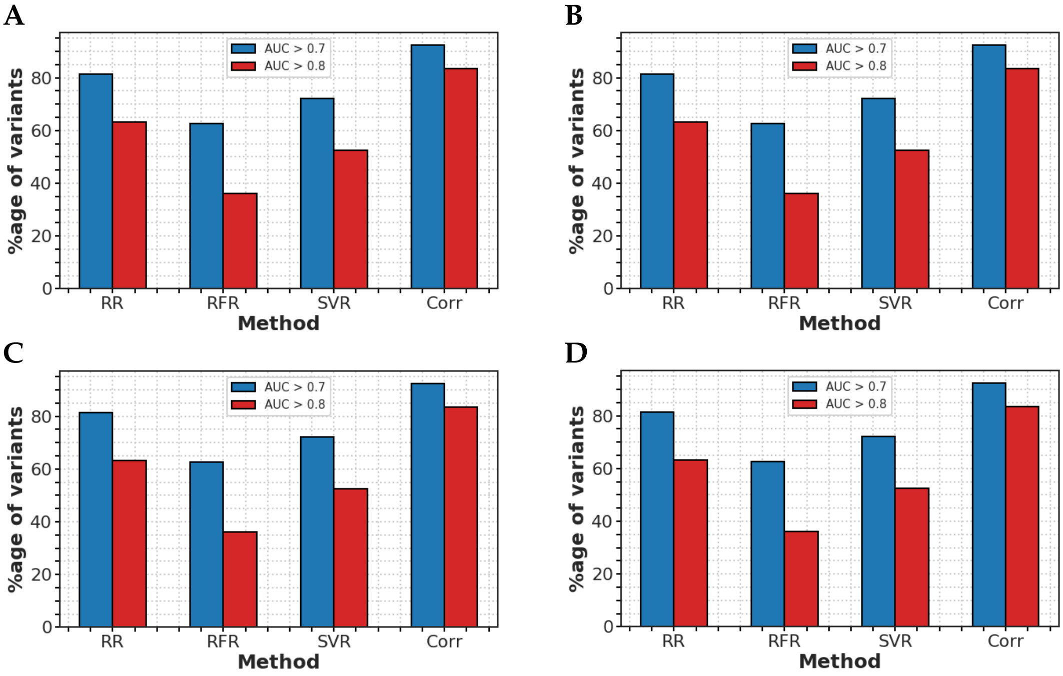

Figure S9: Bar plots show the proportion of variants with trans-eQTL target prediction AUROC > 0.7 (blue) and > 0.8 (red) for random forest regression (RFR), support vector regression (SVR), ridge regression (RR), and univariate correlation (Corr). (**A**) DREAM Network 2, (**B**) DREAM Network 3, (**C**) DREAM Network 4, (**D**) DREAM Network 5.

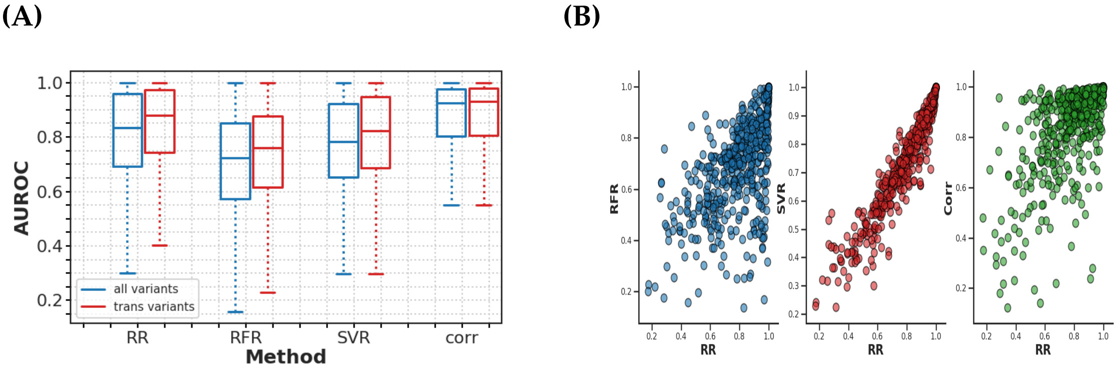

Figure S10: Trans-eQTL target prediction performance on DREAM5 simulated data. **(A)** Boxplots show the distribution of AUROC values for all variants (blue) and for trans-acting-only variants (red) for random forest regression (RFR), support vector regression (SVR), ridge regression (RR), and univariate correlation (Corr). **(B)** Scatter plots show AUROC values of classification methods RFR, SVR, and Corr vs RR for all variants. The data shown are for **DREAM Network 2**.

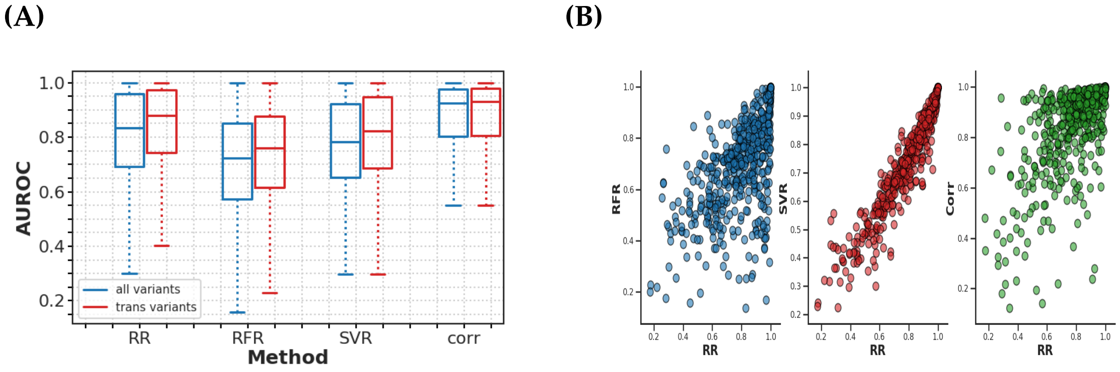

Figure S11: Trans-eQTL target prediction performance on DREAM5 simulated data. **(A)** Boxplots show the distribution of AUROC values for all variants (blue) and for trans-acting-only variants (red) for random forest regression (RFR), support vector regression (SVR), ridge regression (RR), and univariate correlation (Corr). **(B)** Scatter plots show AUROC values of classification methods RFR, SVR, and Corr vs RR for all variants. The data shown are for **DREAM Network 3**.

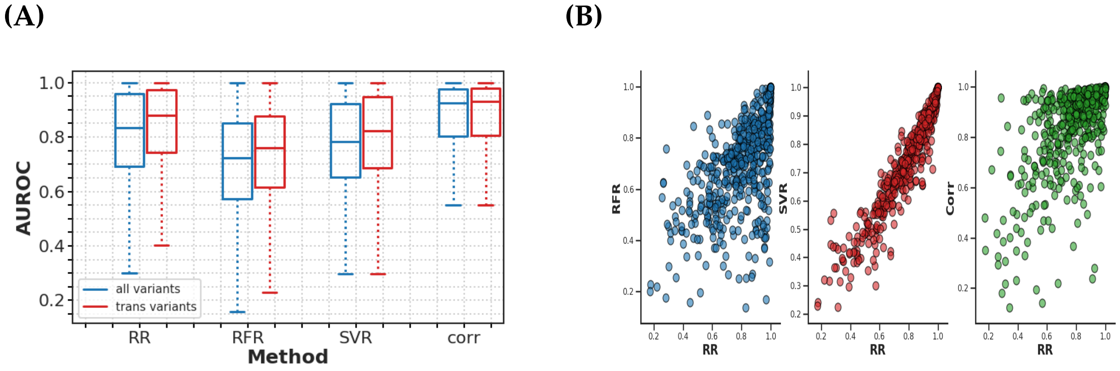

Figure S12: Trans-eQTL target prediction performance on DREAM5 simulated data. **(A)** Boxplots show the distribution of AUROC values for all variants (blue) and for trans-acting-only variants (red) for random forest regression (RFR), support vector regression (SVR), ridge regression (RR), and univariate correlation (Corr). **(B)** Scatter plots show AUROC values of classification methods RFR, SVR, and Corr vs RR for all variants. The data shown are for **DREAM Network 4**.

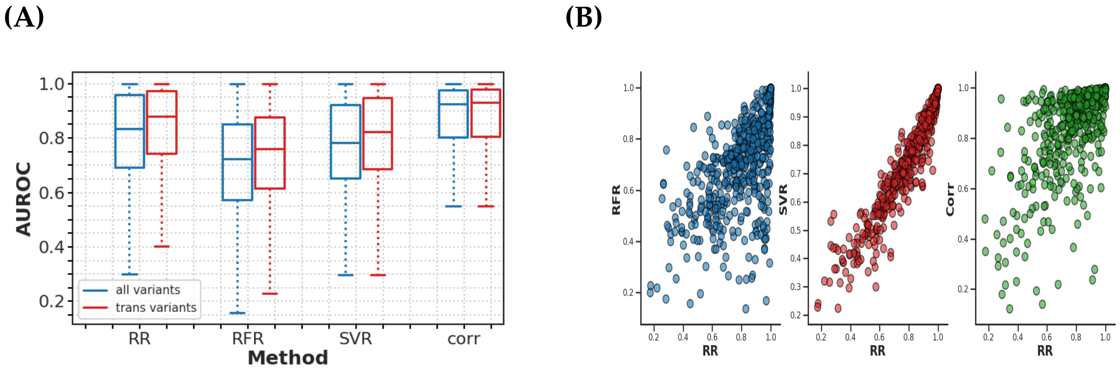

Figure S13: Trans-eQTL target prediction performance on DREAM5 simulated data. **(A)** Boxplots show the distribution of AUROC values for all variants (blue) and for trans-acting-only variants (red) for random forest regression (RFR), support vector regression (SVR), ridge regression (RR), and univariate correlation (Corr). **(B)** Scatter plots show AUROC values of classification methods RFR, SVR, and Corr vs RR for all variants. The data shown are for **DREAM Network 5**.

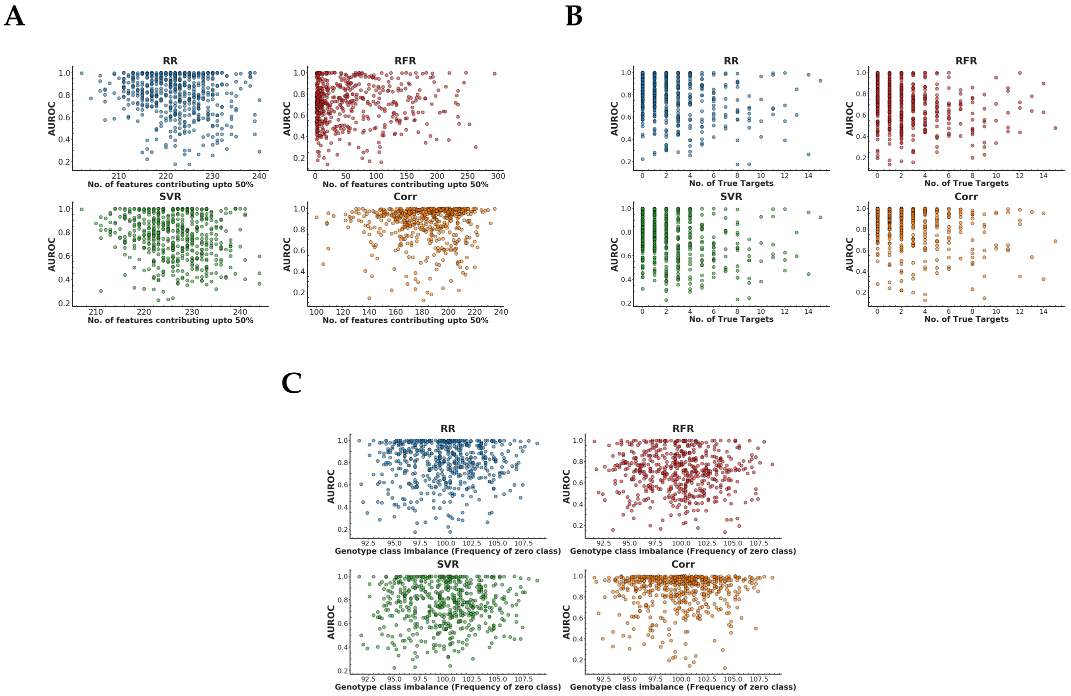

Figure S14: Scatter plots of trans-eQTL target prediction performance (AUROC) on DREAM5 simulated data against the number of selected model features (**A**), the number of true trans-eQTL targets in the ground-truth network (**B**), and the genotype class balance (frequency of the zero class) (**C**), for random forest regression (RFR), support vector regression (SVR), ridge regression (RR), and univariate correlation/naive Bayes (NB). The data shown are for **DREAM Network 2**.

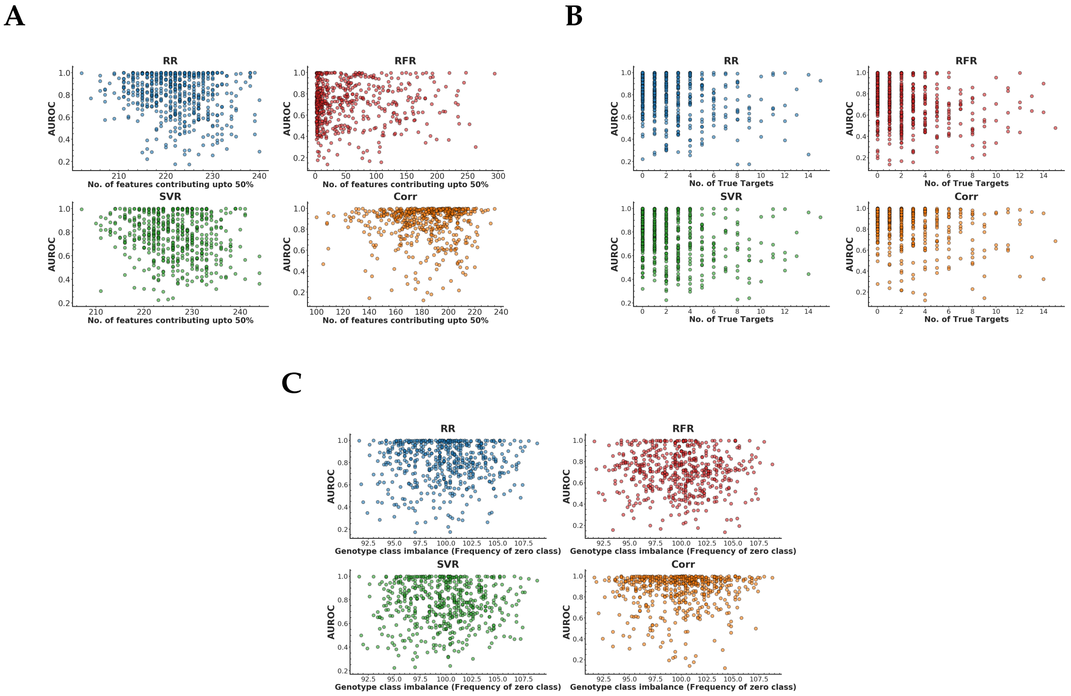

Figure S15: Scatter plots of trans-eQTL target prediction performance (AUROC) on DREAM5 simulated data against the number of selected model features (**A**), the number of true trans-eQTL targets in the ground-truth network (**B**), and the genotype class balance (frequency of the zero class) (**C**), for random forest regression (RFR), support vector regression (SVR), ridge regression (RR), and univariate correlation/naive Bayes (NB). The data shown are for **DREAM Network 3**.

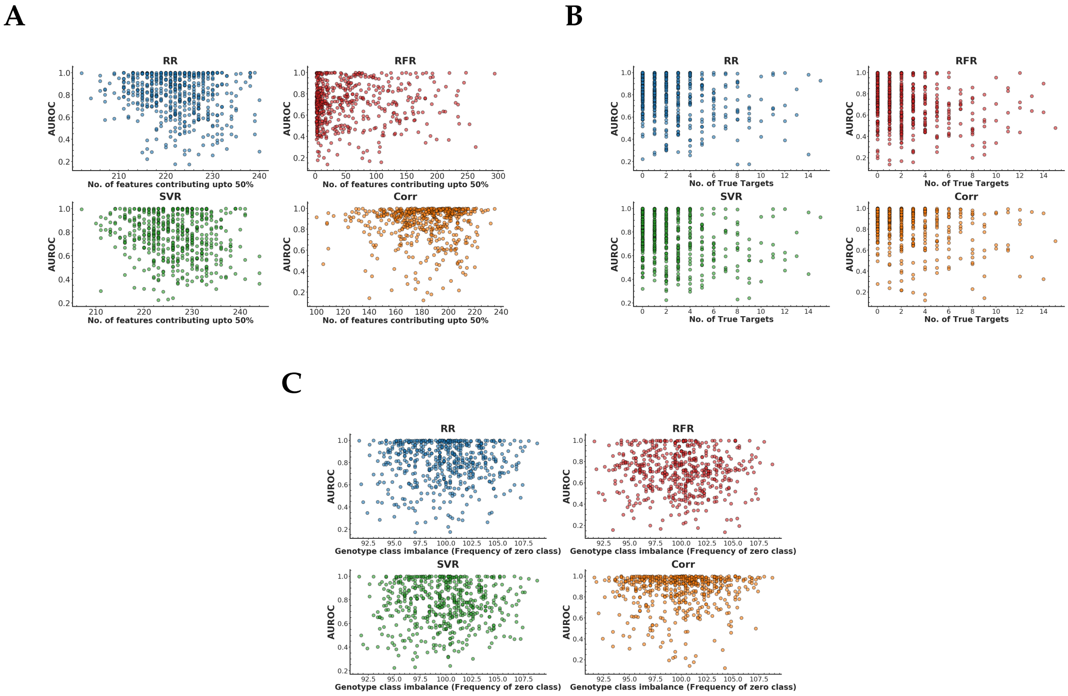

Figure S16: Scatter plots of trans-eQTL target prediction performance (AUROC) on DREAM5 simulated data against the number of selected model features (**A**), the number of true trans-eQTL targets in the ground-truth network (**B**), and the genotype class balance (frequency of the zero class) (**C**), for random forest regression (RFR), support vector regression (SVR), ridge regression (RR), and univariate correlation/naive Bayes (NB). The data shown are for **DREAM Network 4**.

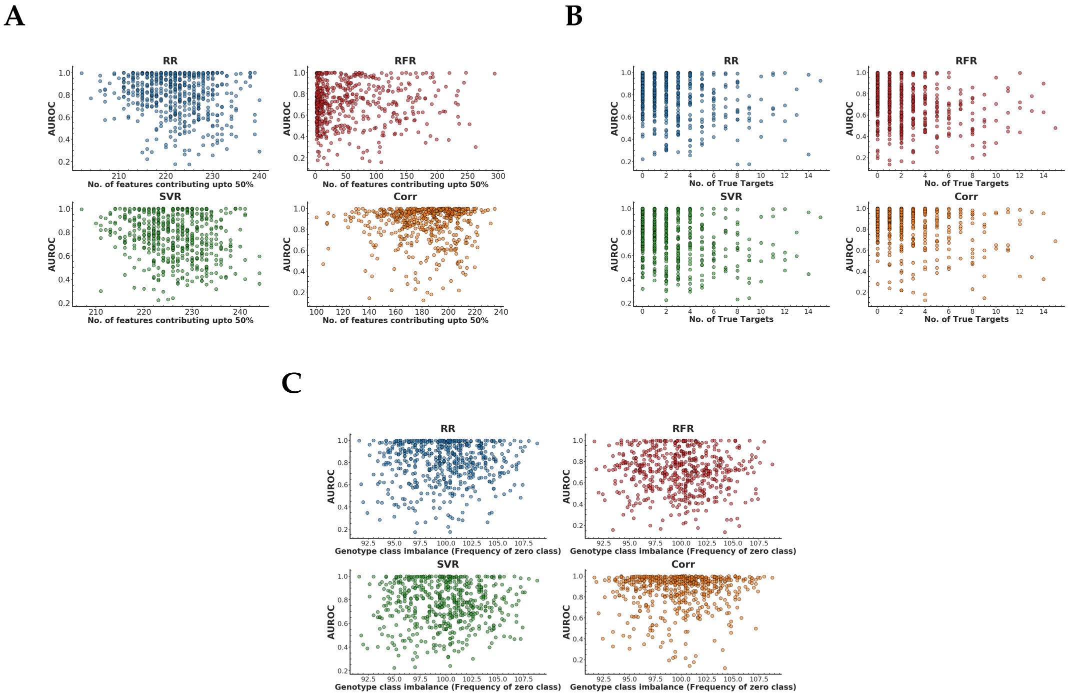

Figure S17: Scatter plots of trans-eQTL target prediction performance (AUROC) on DREAM5 simulated data against the number of selected model features (**A**), the number of true trans-eQTL targets in the ground-truth network (**B**), and the genotype class balance (frequency of the zero class) (**C**), for random forest regression (RFR), support vector regression (SVR), ridge regression (RR), and univariate correlation/naive Bayes (NB). The data shown are for **DREAM Network 5**.

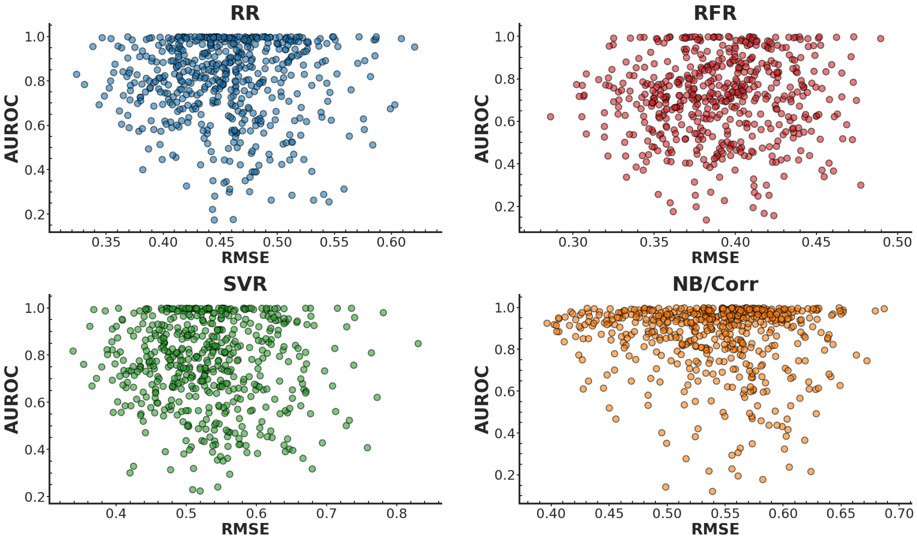

Figure S18: Scatter plots show trans-eQTL target prediction performance (AUROC) vs genotype prediction performance (RMSE) on DREAM5 simulated data for all genetic variants for random forest regression (RFR), support vector regression (SVR), ridge regression (RR), and univariate correlation/naive Bayes (NB/Corr). The data shown are for **DREAM Network 2**.

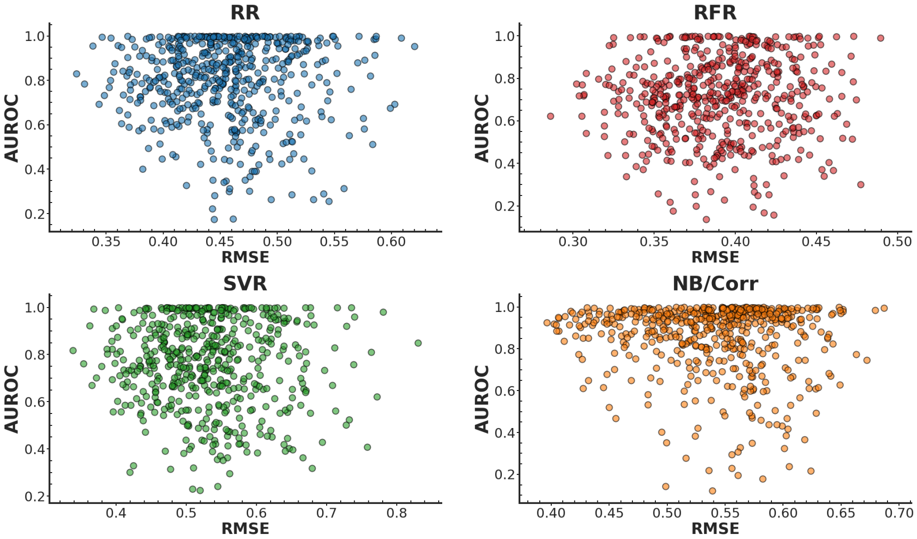

Figure S19: Scatter plots show trans-eQTL target prediction performance (AUROC) vs genotype prediction performance (RMSE) on DREAM5 simulated data for all genetic variants for random forest regression (RFR), support vector regression (SVR), ridge regression (RR), and univariate correlation/naive Bayes (NB/Corr). The data shown are for **DREAM Network 3**.

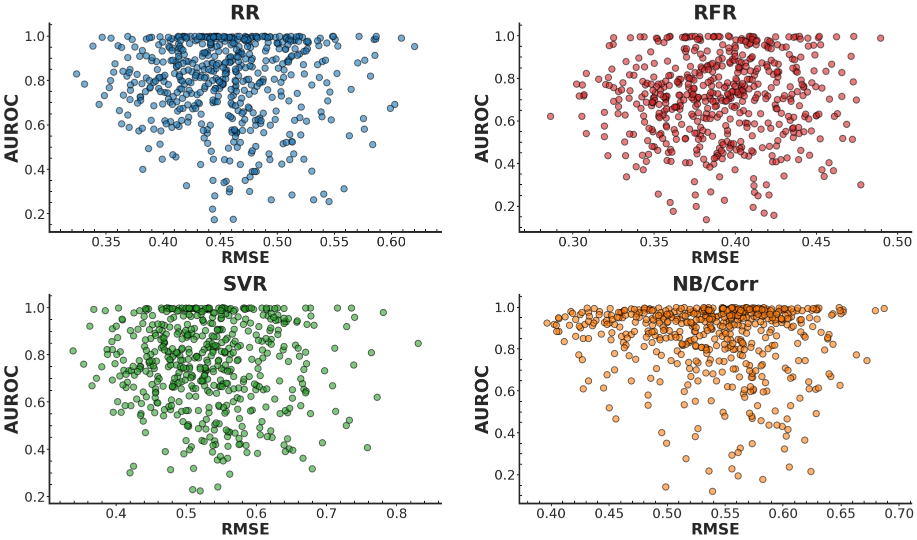

Figure S20: Scatter plots show trans-eQTL target prediction performance (AUROC) vs genotype prediction performance (RMSE) on DREAM5 simulated data for all genetic variants for random forest regression (RFR), support vector regression (SVR), ridge regression (RR), and univariate correlation/naive Bayes (NB/Corr). The data shown are for **DREAM Network 4**.

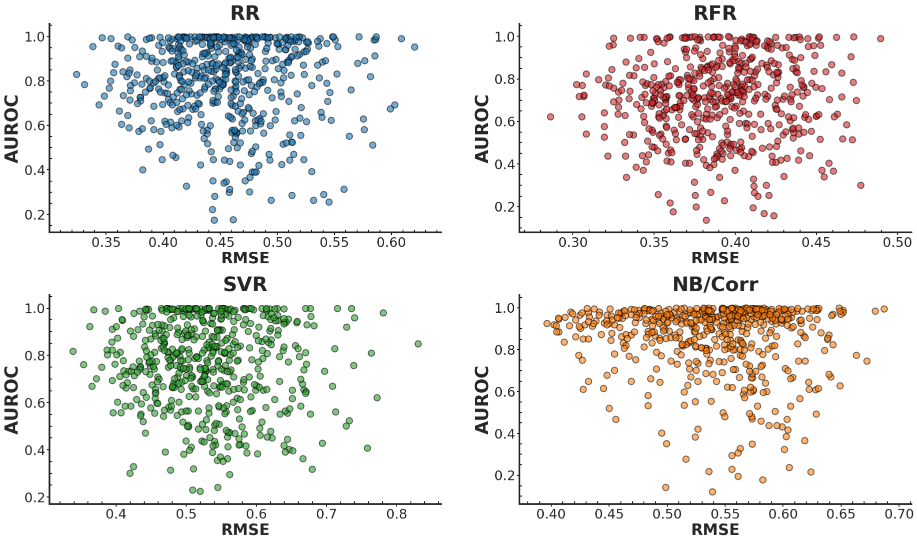

Figure S21: Scatter plots show trans-eQTL target prediction performance (AUROC) vs genotype prediction performance (RMSE) on DREAM5 simulated data for all genetic variants for random forest regression (RFR), support vector regression (SVR), ridge regression (RR), and univariate correlation/naive Bayes (NB/Corr). The data shown are for **DREAM Network 5**.

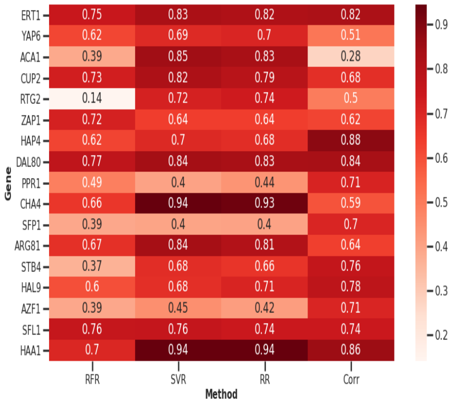

Figure S22: AUROC values for genes where at least one of the four methods (RFR, SVR, RR, Corr) gives AUROC above 0.7.

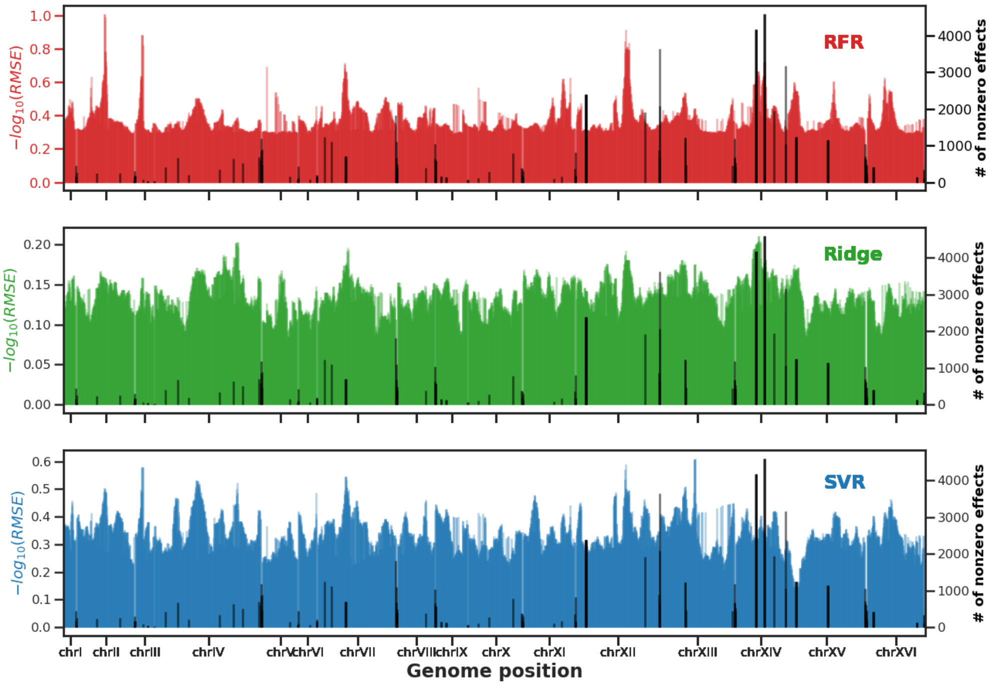

Figure S23: Expression hotspot maps showing the negative log transformed RMSE values vs genome position for 2884 SNPs in the yeast genome, for random forest (RF, top), ridge regression (Ridge, middle), and support vector regression (SVR, bottom). Genes on the same chromosome were excluded as predictors for each SNP. Secondary axis on right shows number of non-zero effects of trans-regulatory hotspot variants from Albert et al. (2018)[5].

[5] Albert, F. W. et al. (2018). Genetics of trans-regulatory variation in gene expression. Elife, 7, e35471