# ADDO: a comprehensive toolkit to detect, classify and visualise additive and non-additive Quantitative Trait Loci

**Leilei Cui[1†], Bin Yang[1†], Nikolas Pontikos[2,3,4], Richard Mott[2*] and Lusheng Huang[1*]**

[1]State Key Laboratory for Pig Genetic Improvement and Production Technology, Jiangxi Agricultural University, Nanchang, P. R. China, [2]UCL Genetics Institute, University College London, London, United Kingdom, [3]UCL Institute of Ophthalmology, University College London, London, United Kingdom and [4]Department of Genetics, Moorfields Eye Hospital, London, United Kingdom

†These authors contributed equally to this work

*To whom correspondence should be addressed

## Abstract

**Motivation**: During the past decade, genome-wide association studies (GWAS) have been used to map quantitative trait loci (QTLs) underlying complex traits. However, most GWAS focus on additive genetic effects while ignoring non-additive effects, on the assumption that most QTL act additively. Consequently, QTLs driven by dominance and other non-additive effects could be overlooked.

**Results**: We developed ADDO, a highly-efficient tool to detect, classify and visualise quantitative trait loci (QTLs) with additive and non-additive effects. ADDO implements a mixed-model transformation to control for population structure and unequal relatedness that accounts for both additive and dominant genetic covariance among individuals, and decomposes single nucleotide polymorphism (SNP) effects as either additive, partial dominant, dominant and over-dominant. A matrix multiplication approach is used to accelerate the computation: a genome scan on 13 million markers from 900 individuals takes about 5 hours with 10 CPUs. Analysis of simulated data confirms ADDO's performance on traits with different additive and dominance genetic variance components. We showed two real examples in outbred rat where ADDO identified significant dominant QTL that were not detectable by an additive model. ADDO provides a systematic pipeline

to characterize additive and non-additive QTL in whole genome sequence data, which complements current mainstream GWAS software for additive genetic effects.

**Availability and implementation**: ADDO is customizable and convenient to install and provides extensive analytics and visualizations. The package is freely available online at https://github.com/LeileiCui/ADDO.

**Contact**: r.mott@ucl.ac.uk and lushenghuang@hotmail.com

**Supplementary information**: Supplementary data are available at *Bioinformatics* online.

## 1 Introduction

Non-additive genetic effects model the interactions between alleles at a locus (Visscher *et al*., 2008), and are distinct from additive genetic effects, where a heterozygote's effect on a phenotype lies midway between those of the two homozygotes. They are classified into partial dominance (heterozygote deviates from the average of the homozygotes but does not exceed either value), complete dominance (heterozygote similar to one homozygote) and overdominance (substantially outside the range of two homozygotes) (Ungerer, 2004). Dominance and overdominance are likely causes of the important phenomenon of heterosis (Bruce, 1910; Jones, 1917; Shull, 1908). A careful characterization of genetic effects is important in human genetics as well, in order to understand the complex mechanism of many human quantitative traits and diseases. It is also essential in order to understand how natural selection and sexually antagonistic selection operate in wild species (Barson *et al*., 2015), and for genomic selection in livestock (Akanno, *et al*., 2018; Wellmann and Bennewitz, 2012), e.g. to guide mate allocation in animal crossbreeding (Wellmann and Bennewitz, 2012).

Most GWAS software employ mixed linear models (Kang *et al*., 2010; Yu, *et al*., 2006; Kang, *et al*., 2008; Ning, *et al*., 2018; Zhou and Matthew, 2012), which utilise a genetic relationship matrix (or kinship matrix) to model the varying degrees of relationships among individuals. While these mixed models - often in combination with principle components as fixed effects - control population stratification, produce well calibrated *P*-values, they usually ignore dominance and other non-additive effects. The motivations behind the assumption of additivity are probably that (i) most causal variants are thought to act additively, so the extra degrees of freedom required to test for non-additive effects would be expected to reduce power slightly, and (ii) allele frequencies are often too low to estimate the phenotypic effect of the rarer of the two homozygotes reliably, unless sample sizes are large, so there is insufficient information to test for non-additivity.

Nonetheless GWAS based on additivity miss the opportunity to identify non-additive QTL, and which are of particular importance in animal and crop genetics. Moreover, if one's interest is in genetic architecture and mechanism rather than discovery, it is necessary to understand each QTL's mode of action. Although, studies in humans showed

that dominance genetic variation has limited contribution to the missing heritability for complex traits (Zhu *et al.,* 2015), such studies are still rare, and investigations on wider range of phenotypes and species is needed to clarify non-additive genetic architecture of complex traits. Moreover, as far as we know, there is very few program explicitly examined a QTL's mode of action in term of additive and dominant effects.

Here, we describe an R package, ADDO, for the efficient detection and classification of non-additive QTL, ADDO implements a linear mixed model to control population structure that explicitly models non-additive effects (Kang *et al*., 2008), and which utilizes large matrix operations to speed up computation (Shabalin, 2012). ADDO also provides versatile functions to classify and visualize non-linear association results.

## 2 Algorithm

### 2.1 Linear mixed model for non-additive genetic effects

ADDO uses well-established theory for modeling non-additive effects. A general linear mixed model to investigate the association between a single SNP with a given trait can be expressed as:

$$\boldsymbol{y} = \boldsymbol{M\beta} + \boldsymbol{u} + \boldsymbol{e}$$

where $\boldsymbol{y}$ is a vector of phenotypic residuals that have been corrected for environmental fixed effects and other covariates. $\boldsymbol{M}$ is the matrix encoding the genotype effects of a given SNP, $\boldsymbol{\beta}$ is a vector of genotype effects depending on the coding of $\boldsymbol{M}$. $\boldsymbol{u}$ and $\boldsymbol{e}$ are vectors of random genetic background and residual random effects, with covariance matrices being $\sigma_g^2\boldsymbol{K}$ and $\sigma_e^2\boldsymbol{I}$, respectively, where $\boldsymbol{K}$ is the kinship matrix and $I$ is the identity matrix. For additive effects, the three genotypes of an SNP (AA, AB and BB), are encoded by a vector (0, 1, 2). To estimate and classify the dominance effects of one locus, we use two essentially equivalent codings:

$$Additive - Dominant\ Matrix,\ \boldsymbol{M_{AD}}:$$

$$\begin{matrix} AA \\ AB \\ BB \end{matrix}\begin{pmatrix} 0 & 0 \\ 1 & 1 \\ 2 & 0 \end{pmatrix}$$

$$Indicative\ Matrix,\ \boldsymbol{M_I}:$$

$$\begin{matrix} AA \\ AB \\ BB \end{matrix}\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Thus, $\boldsymbol{M_{AD}}$ augments the one-column matrix representing an additive effect with a second column that models the deviation of the heterozygote from its expected value under the additive model. $\boldsymbol{M_I}$ models each genotype with a separate effect. When $\boldsymbol{M_{AD}}$ is augmented with an intercept term (i.e. a column of 1's), it is an invertible transformation of $\boldsymbol{M_I}$. We use these equivalent models to explore different aspects of non-additivity, as described below.

We model the phenotypic covariance matrix of $\boldsymbol{y}$ as $\boldsymbol{V} = \sigma_a^2\boldsymbol{K_a} + \sigma_d^2\boldsymbol{K_d} + \sigma_e^2\boldsymbol{I}$, where $\sigma_a^2$, $\sigma_d^2$ and $\sigma_e^2$ are the additive, dominance and residual variance components, as estimated by GREML, e.g. GCTA (Yang *et al*., 2011). The inclusion of dominance variance is logically consistent with our objective of modeling non-additive effects and moreover improves the calibration of GWAS non-additive *P*-values. $\boldsymbol{K_a}$ and $\boldsymbol{K_d}$ are the additive and dominance kinship matrices among individuals, calculated using the standard

definition in GCTA, and $\boldsymbol{I}$ is the identity matrix,

$$K_{a(ij)} = \frac{1}{m}\sum_k \frac{(x_{a(ik)} - 2p_k)(x_{a(jk)} - 2p_k)}{2p_k(1-p_k)}$$

$$K_{d(ij)} = \frac{1}{m}\sum_k \frac{(x_{d(ik)} - 2p_k^2)(x_{d(jk)} - 2p_k^2)}{4p_k^2(1-p_k)^2}$$

where $K_{a(ij)}$ and $K_{d(ij)}$ are the elements of $\boldsymbol{K_a}$ and $\boldsymbol{K_d}$ between individuals $i$ and $j$; $x_{a(ik)}$ and $x_{d(ik)}$ are the additive and dominant genotype coding of individual $i$ in SNP $k$, which $x_{a(ik)} = 0, 1\ or\ 2$ and $x_{d(ik)} = 0, 2p\ or\ (4p-2)$ for three genotypic classes AA, AB and BB; $m$ is the total number of SNPs and $p_k$ is the frequency of allele B in SNP $k$.

$\boldsymbol{V}$ is factorized into its matrix square root $\boldsymbol{A}$ through eigenvalue decomposition using the R function *eigen()* (R Development Core Team, 2013):

$$\boldsymbol{V = UDU^{-1} = (UA^{1/2}U^{-1})^2 = A^2}$$

where $\boldsymbol{U}$ is the matrix of eigenvectors and $\boldsymbol{D}$ is the diagonal matrix with the eigenvalues. We next transform the mixed model by multiplying both side of the equation with inverse matrix of $\boldsymbol{A}$:

$$\boldsymbol{A^{-1}y = (A^{-1}M)\beta + A^{-1}(u+e)}$$

Where $\boldsymbol{M}$ is either $\boldsymbol{M_{AD}}$ or $\boldsymbol{M_I}$, after the transformation, the variance matrix of the model residual term, $\boldsymbol{A^{-1}(u+e)}$ is the identity matrix $\boldsymbol{I}$, and the vector of SNP effects $\boldsymbol{\beta}$ could be estimated with an ordinary linear model using the R function *lm()* and the statistical significance (-$\log_{10}$ *P-value*) were calculated by the analysis of variance (ANOVA) comparing the fit of the transformed mixed model to that of the transformed null model.

**2.2 Fast detection for significant loci through matrix operations**

In order to accelerate the ADDO package, we apply a matrix operation strategy to speed up the genome wide testing for variants set that are significant associated with the target trait. This is achieved through replacing the standard linear regression procedure by a large matrix multiplication with standardized and orthogonalized variables as in (Shabalin, 2012). The statistic $\boldsymbol{R^2}$ (the fitting sum of square) is estimated as follows:

(1) Transform the residual vector $\boldsymbol{y}$ and two genotype vectors $\boldsymbol{M_A}$ and $\boldsymbol{M_D}$, which are the first and second columns of $\boldsymbol{M_{AD}}$, to correct the population stratification effect

$$\boldsymbol{y_T = A^{-1}y,\ M_{AT} = A^{-1}M_A,}$$

$$\boldsymbol{M_{DT} = A^{-1}M_D}$$

(2) Orthogonalize $\boldsymbol{M_{DT}}$ with respect to $\boldsymbol{M_{AT}}$ for each locus, in what follows, $\langle \boldsymbol{u,v} \rangle$ denotes the inner product between vectors $\boldsymbol{u,v}$.

$$\widetilde{\boldsymbol{M_{DT}}} = \boldsymbol{M_{DT} - \langle M_{DT}, M_{AT} \rangle M_{AT}}$$

(3) Standardize $\boldsymbol{y_T}$, $\boldsymbol{M_{AT}}$ and $\widetilde{\boldsymbol{M_{DT}}}$

(4) Estimate the test statistic $\boldsymbol{R^2}$ through large matrix operations

$$\boldsymbol{R^2 = \langle y_T, M_{AT} \rangle^2 + \left\langle y_T, \widetilde{M_{DT}} \right\rangle^2}$$

(5) Calculate the F-test from the test statistics $\boldsymbol{R^2}$ and the statistical significance (-$log_{10}$ *P-value*) using R function *pf* (Team, 2013).

$$\boldsymbol{F = \frac{(n-3)R^2}{2(1-R^2)}}$$

**2.3 Classification of QTLs by ratio of dominance and additive effects**

Next, we identify suggestive and genome-wide significant QTLs using Bonferroni correction for M SNPs, with $-log_{10}\left(\frac{1}{M}\right)$ and $-log_{10}\left(\frac{0.05}{M}\right)$ as the default significance thresholds, respectively (the former corresponding to one expected false positive per genome scan, and the latter to 5% genome-wide significance). To characterize the contribution of additive and dominance effects of significant QTLs, we refit the same model but using different parameterisations. We extract additive ($\beta_{Add}$) and dominance ($\beta_{Dom}$) effects and respective standard errors $se(\beta_{Add})$ *and* $se(\beta_{Dom})$ using the Additive – Dominance effect incidence matrix, $\boldsymbol{M_{AD}}$:

$$\boldsymbol{A^{-1}y = (A^{-1}M_{AD})\beta + A^{-1}(u + e)}$$

(Add-Dom Model)

Then, we calculate two t-statistics $t_{Add}$ and $t_{Dom}$, corresponding to the standardized QTL additive and dominance effects, respectively:

$$t_{Add} = \frac{\beta_{Add}}{se(\beta_{Add})} \qquad t_{Dom} = \frac{\beta_{Dom}}{se(\beta_{Dom})}$$

We categorize QTL using the ratio of the two t statistics: QTL with $|t_{Dom}/t_{Add}| < 0.2$, $0.2 < |t_{Dom}/t_{Add}| < 0.8$, $0.8 < |t_{Dom}/t_{Add}| < 1.2$, and $1.2 < |t_{Dom}/t_{Add}|$ are classified as additive, partial dominance, dominance and overdominance QTL, respectively (Supplementary Figure S1).

### 2.4 Verification of over-dominant QTLs

To further determine and verify overdominant/heterotic QTL, we refit the QTL using the indicative coding matrix $\boldsymbol{M_I}$ to estimate the genetic effect of each genotype:

$$\boldsymbol{A^{-1}y = (A^{-1}M_I)\beta + A^{-1}(Zu + e)}$$

(Heterotic Model)

For each locus, we extract the effects of three genotypes $\beta_{AA}$, $\beta_{AB}$ and $\beta_{BB}$, and calculate two T-statistics to measure the deviation between the effect of heterozygote (AB) and that of two homozygotes (AA and BB),

$$t_1 = t(AB - AA) = \frac{\beta_{AB} - \beta_{AA}}{se(\beta_{AB} - \beta_{AA})}$$

$$t_2 = t(AB - BB) = \frac{\beta_{AB} - \beta_{BB}}{se(\beta_{AB} - \beta_{BB})}$$

In order to classify a given QTL and to determine its statistical significance, we combine all T-statistics to generate two vectors $\boldsymbol{t_1}$ and $\boldsymbol{t_2}$, and chose the minimum of the absolute value of those two T-statistics as $\boldsymbol{t_M}$ to estimate the statistical significance based on the multivariate normal (MVN) distribution function *pmvnorm* from R package *mvtnorm* (Genz and Bretz, 2010),

$$P_{(\boldsymbol{t_1},\boldsymbol{t_2})} = \frac{a_1 + a_2}{A_1 + A_2}$$

$$= \frac{\int_{\boldsymbol{t_M}}^{\infty}\int_{\boldsymbol{t_M}}^{\infty} D(\mu_1,\mu_2,\rho)d\boldsymbol{t_1}d\boldsymbol{t_2} + \int_{-\infty}^{-\boldsymbol{t_M}}\int_{-\infty}^{-\boldsymbol{t_M}} D(\mu_1,\mu_2,\rho)d\boldsymbol{t_1}d\boldsymbol{t_2}}{\int_{0}^{\infty}\int_{0}^{\infty} D(\mu_1,\mu_2,\rho)d\boldsymbol{t_1}d\boldsymbol{t_2} + \int_{-\infty}^{0}\int_{-\infty}^{0} D(\mu_1,\mu_2,\rho)d\boldsymbol{t_1}d\boldsymbol{t_2}}$$

where $a_1$ is the area from MVN distribution function defined by $\boldsymbol{t_1},\boldsymbol{t_2} > \boldsymbol{t_M}$, $a_2$ is the area where $\boldsymbol{t_1},\boldsymbol{t_2} < -\boldsymbol{t_M}$, $A_1$ is the area where $\boldsymbol{t_1}$, $\boldsymbol{t_2} > 0$, $A_2$ is the area where $\boldsymbol{t_1},\boldsymbol{t_2} < 0$, $\mu_1$ and $\mu_2$ are the mean values of $\boldsymbol{t_1}$ and $\boldsymbol{t_2}$, and $\rho$ is the correlation coefficient of $\boldsymbol{t_1}$ and $\boldsymbol{t_2}$ (Supplementary Figure S2).

## 3 Results

### 3.1 Comparison of ADDO with GenABEL, EMMAX, GEMMA and GCTA

To verify the reliability of the algorithm implemented in ADDO, we compared the

SNP-trait association *P*-values calculated by ADDO with four other GWAS software, GenABEL (Aulchenko *et al*., 2007), EMMAX (Kang, *et al*., 2010), GEMMA (Zhou and Matthew, 2012), which all implemented the additive genotypic model. The example dataset we used is from a Rat Heterogeneous Stock (HS rats) with 160 traits and 244,876 SNP genotypes of 1,407 individuals (Baud *et al*., 2013). For all SNPs tested, the *P*-values from ADDO were highly correlated with those obtained the other programs (Fig 1), suggesting statistics obtained from ADDO are reliable. The inflation factor (calculated by dividing the observed mediation chi-square statistics with expected median chi-square statistics) was 1.007, suggesting the population structure can be properly controlled by ADDO.

### 3.2 Simulation study

We performed a simulation study to test the power of our model on phenotypes controlled by QTLs with different proportions of additive and dominant variances components. The phenotypes were simulated based on the observed genotypes of 1,407 individuals from the rat heterogeneous stock (Baud, et al., 2013). Firstly, we randomly selected 1,000 SNPs across the genome that are in weak linkage disequilibrium among each other ($r^2$<0.1) and with high minor allele frequency (MAF: 0.4-0.5) using Plink (Purcell, et al., 2007). Then 50 SNPs were randomly selected from these 1000 SNPs to become QTLs, and at which we simulated five classes of phenotypes, namely (1) solely additive genetic variance, i.e. $V_a = 1\%$; (2) greater additive and smaller dominance genetic variance, $V_a = 1\%$ and $V_d = 0.1\%$, respectively; (3) equal additive and dominance genetic variance, $V_a = 0.5\%$ and $V_d = 0.5\%$, respectively; (4)

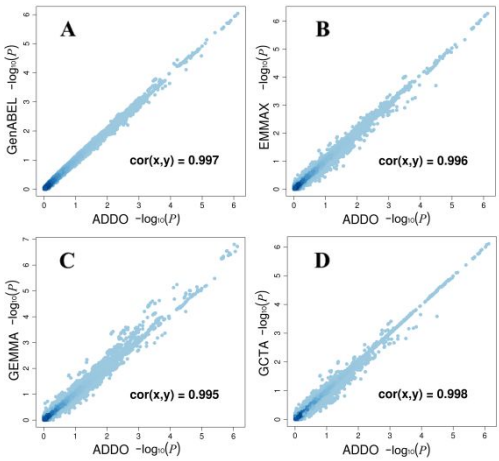


**Fig 1.** Comparison of *P*-values of ADDO with those obtained by GenABEL, EMMAX, GEMMA and GCTA based on an anxiety related behavioural trait and 244,876 SNPs of 1,340 individuals from a rat heterogeneous stock.

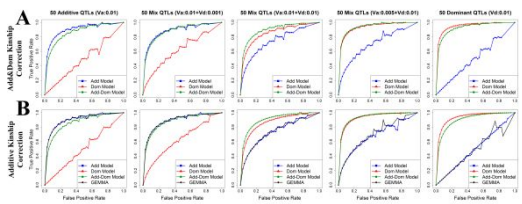


**Fig 2. Evaluation and comparison of GEMMA and different models implemented in ADDO on simulated phenotype using ROC curves.** Five set of simulated phenotypes with different proportions of additive and dominance genetic components were represented in five columns. (A) Models that consider only additive polygenic effects, $\boldsymbol{K_a}$ (B) Models that correct for both additive and dominance polygenic effects, $\boldsymbol{K_a}$ and $\boldsymbol{K_d}$. Each column stands for one type of variance combination to simulate phenotype. The colors of lines stand for different genotypic coding matrix used by ADDO.

smaller additive and greater dominance genetic variance, $V_a = 0.5\%$ and $V_d = 1\%$,

respectively; (5) only dominance genetic variance, $V_d = 1\%$. Note that the numbers refer to the % target phenotypic variance explained by each simulated QTL, not to the total genetic variance, which will be close to the sum of the individual QTL effects. Next, the corresponding additive effect $a$ and dominance effect $d$ of each QTL were calculated based on $V_a = 2p(1-p)[a+(1-2p)d]^2$ and $V_d = [2p(1-p)d]^2$, where $p$ is the minor allele frequency of the specified QTL. Environmental effects were simulated to follow the standard normal distribution with variance adjusted to ensure the total genetic variance was around 33.3%, a typical value for this population. We generated the simulated phenotypes by summing up additive and/or dominance effects of all 50 QTLs, and the environmental effects.

Then we evaluated the performance of GEMMA and the genotype coding matrices implemented in ADDO through 1000 simulations using receiver operating characteristic (ROC) curves (Fig 2 and Table 1). We compared the three different genotype coding matrices ($M_A$, $M_D$ or $M_{AD}$) to fit SNP effects and two different strategies to control polygenic effects (using either just the additive kinship matrix or the additive + dominance kinship matrix) in ADDO in term of areas under the ROC curve (AUC) (Table 1). To draw the ROC curve, we set a series of thresholds of -$\log_{10}$ *P*-value, from 0 to 10 incremented by 0.1 to call significant QTLs. The true positive rate (TPR) and false positive rate (FPR) corresponding to these thresholds were calculated as:

$$TPR = \frac{Number\ of\ true\ QTLs\ detected\ as\ significant}{Number\ of\ true\ QTLs}$$

$$FPR = \frac{Number\ of\ false\ QTLs\ detected\ as\ significant}{Number\ of\ false\ QTLs}$$

The analysis shows that performance of additive model implemented in ADDO is approximately equivalent to GEMMA, which implements exact mixed model to test the trait-marker association (Zhou *et al*., 2012). Model that simultaneously account for additive and dominance effects (Add-Dom model) out-performed the additive model when the QTLs contain dominance variance. When the simulated QTLs have equal additive and dominance variance component (i.e. $V_a = V_d = 0.5\%$), the Add-Dom model AUC = 0.896, considerably between than AUC = 0.745 for the additive model. Moreover, the Add-Dom model increased AUC from 0.502 to 0.925 when the QTLs with only dominance variance were simulated (i.e. $V_a = 0$ and $V_d = 1\%$). Moreover, the performance of Add-Dom model was robust even when the

**Table 1**. Comparison of performance of GEMMA and different models implemented in ADDO on simulated QTL with different proportion of additive and dominance variance components according to the AUC of the ROC curve.

| Kinship matrix | Model | Variance components of simulated QTL | | | | |
|---|---|---|---|---|---|---|
| | | Va=1 % | Va=1% Vd=0.1% | Va=0.5% Vd=0.5% | Va=0.5% Vd=0.1% | Vd=1 % |
| ADD + DOM | Add | 0.900 | 0.882 | 0.745 | 0.703 | 0.502 |
| | Dom | 0.493 | 0.619 | 0.854 | 0.933 | 0.947 |
| | Add-Dom | 0.864 | 0.880 | 0.896 | 0.939 | 0.925 |
| ADD | Add | 0.900 | 0.888 | 0.740 | 0.677 | 0.502 |
| | Dom | 0.504 | 0.630 | 0.866 | 0.930 | 0.933 |
| | Add-Dom | 0.861 | 0.880 | 0.903 | 0.937 | 0.902 |
| | GEMMA | 0.902 | 0.886 | 0.736 | 0.697 | 0.498 |

QTLs are simulated with only additive variance (i.e. $V_a = 1\%$ and $V_d = 0$), the AUC only dropped from 0.900 to 0.864. Taken together, these results demonstrated that modelling both additive and dominance effects simultaneously has great potential to map additional QTL that may be missed in GWAS when considering only additive effects. We also observed a slight improvement in the model performance when using both the additive and dominance covariance matrices to control for the polygenic genetic effects, when the QTL are simulated with both additive and dominance effects. For instance, for traits simulated with only dominant QTL, the false positive rate (1%) of Add-Dom model accounting for additive and dominance covariance matrices, is substantially lower than that (8.9%) obtained by Add-Dom model considering only additive covariance matrix at a nominal *P* value threshold of 0.01 (**Table S1**).

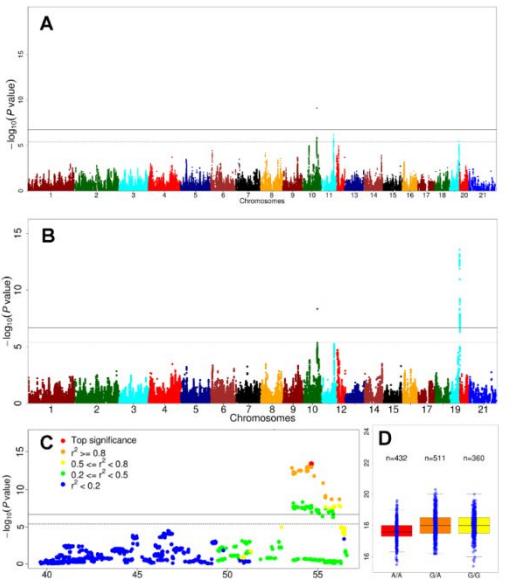


**Fig 3. Comparison of GWAS results on mean corpuscular hemoglobin in a rat heterogeneous stock using GEMMA and ADDO.** (A) Manhattan plot of GWAS by an additive model using GEMMA. (B) Manhattan plot of GWAS implemented by Add-Dom model in ADDO (C) Regional association plot of the significant locus on chromosome 19 identified by ADDO. (D) Boxplot of phenotypic values by the three genotypes of the lead SNP at the significant locus detected by ADDO on chromosome 19.

### 3.3 Results using the real phenotypes

In addition to the simulated phenotypes, we also applied ADDO on real genotype and phenotype data from the rat heterogeneous stock (Baud, et al., 2013), in which around 160 phenotypes relevant to type 2 diabetes, hypertension, multiple sclerosis and anxiety were measured in up to 1,407 individuals. We show results on mean corpuscular hemoglobin (Fig 3) and absolute CD8+ T cell levels (Fig 4) based on the Add-Dom Model implemented in ADDO. For mean corpuscular hemoglobin,

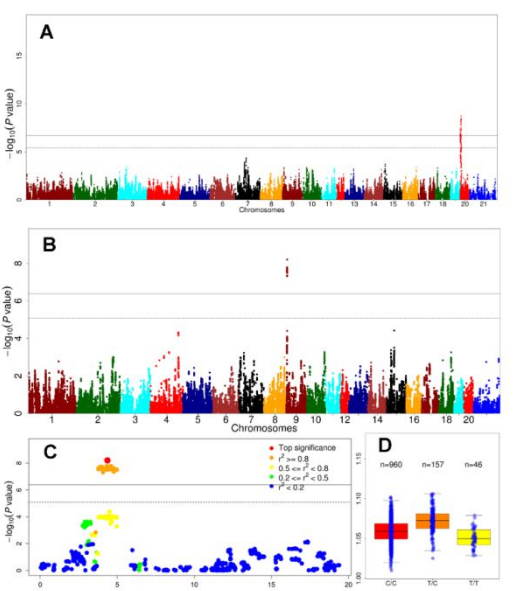


**Fig 4. Comparison of GWAS results on absolute CD8+ T cells in a rat heterogeneous stock using**

**GEMMA and ADDO.** (A) Manhattan plot of GWAS by an additive model using GEMMA. (B) Manhattan plot of GWAS implemented by heterotic model in ADDO. (C) Regional association plot of the significant locus on chromosome 9 identified by ADDO. (D) Boxplot of phenotypic values by the three genotypes of the lead SNP at the significant locus detected by ADDO on chromosome 9.

we identified a highly significant locus on rat chromosome 19 with $-log_{10}$ *P-value* being 13.56, which is much greater than the additive $-log_{10}$ *P-value* of 5.74 implemented in Gemma (Fig 3). For absolute CD8+ T cells, notably, we identified a novel QTL on chromosome 9 ($-\log_{10} P$ value = 8.2) that was also missed by the additive model (Fig 4). The $|t_{Dom}/t_{Add}|$ statistics for the two QTLs were 1.43 and 2.08, thus both QTL were categorized as overdominance QTLs. These two examples support the superior performance of ADDO to detect and classify QTLs with dominance effects over regular additive GWAS software.

## 4 Discussion

In this study, we have shown by analyses of simulated and real data of hematological and immune traits in rat suggested that it is worthwhile to explore non-additive effects in GWAS data from non-human species. The computational time of ADDO depends on the time spend on 1) data loading, 2) inversion of variance-covariance matrix (**V**) of the phenotype, 3) and genomic scan. In current version of ADDO, a genome scan of ADDO on 10 million markers in 2000 individuals takes 23 hours, which is two-fold of the time program lies in the time spending on inversion of **V**, as the computation time will cubically increase with the number of observations used in analysis. Currently, we tested as many as 8000 individuals, and find that the inversion of **V** takes about 10 minutes using *solve()* function in R, and less than 5 minutes with *spdinv()* function from Rfast package (https://rdrr.io/cran/Rfast/) (Supplementary Figure S3). In terms of this analysis, we consider that the ADDO program is able to scale up to analyze data set of 10000-20000 individuals with whole genome sequence marker data. There are several other fast and flexible linear model implementations such as Grid-LMM (Raniel and Lorin, 2019), however, this approach is more powerful when fitting more than two random effects.

Overall, the R package ADDO developed here not only provides a tool to detect additive and dominance QTL, which helps to better understand the genetic effects genomic variants on complex traits. By implementing a matrix multiplication strategy to speed up the computations, the ADDO is also potentially applicable to intermediate molecular data such as transcriptome and proteome measurements of thousands of traits.

### Acknowledgements

We acknowledge Dr. Amelie Baud, Dr. Na Cai, Dr. Michael Scott, Dr. Sanja Franic and Dr. Jack Humphrey for their helpful advice and suggestions regarding algorithm design and software implementation.

*Conflict of Interest*: none declared.

### References

Akanno, E.C., *et al.* (2018) Genome-wide

association scan for heterotic quantitative trait loci in multi-breed and crossbred beef cattle, *Genet Sel Evol*, 50.

Aulchenko, Y.S. *et al.* (2007) GenABEL: an R library for genome-wide association analysis, *Bioinformatics,* 23, 1294-1296.

Barson, N.J. *et al.* (2015) Sex-dependent dominance at a single locus maintains variation in age at maturity in salmon, *Nature,* 528, 405.

Baud, A. *et al.* (2013) Combined sequence-based and genetic mapping analysis of complex traits in outbred rats, *Nat Genet*, 45, 767-775.

Bruce, A.B. (1910) The mendelian theory of heredity and the augmentation of vigor, *Science*, 32, 627-628.

Daniel, E. and Lorin, C., (2019) Fast and flexible linear mixed models for genome-wide genetics, *Plos Genet*, **15**, 2.

Genz, A. and Bretz, F. (2010) Computation of Multivariate Normal and t Probabilities, *J Stat Softw*, 33, 1641.

Kang, H. *et al.* (2010) Variance component model to account for sample structure in genome-wide association studies., *Nat Genet*, 42, 348-354.

Yu, J. *et al.* (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness, *Nat Genet*, 38, 203-208.

Jones, D.F. (1917) Dominance of Linked Factors as a Means of Accounting for Heterosis, P NATL ACAD SCI USA, 2, 466-479.

Kang, H. *et al.* (2008) Efficient control of population structure in model organism association mapping, *Genetics*, 178, 1709.

Ning, C. *et al.* (2018) A Rapid Epistatic Mixed-model Association Analysis by Linear Retransformations of Genomic Estimated Values, *Bioinformatics*, 34.

Purcell, S. *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses, *The American Journal of Human Genetics*, 81, 559-575.

Shabalin, A.A. (2012) Matrix eQTL: ultra fast eQTL analysis via large matrix operations, *Bioinformatics*, 28, 1353-1358.

Shull, G.H. (1908) The composition of a field of maize, *Reports of the American Breeders Association*, 296-301.

R Development Core Team (2013) R: A language and environment for statistical computing.

Ungerer, M.C. (2004) A Primer of Ecological Genetics.

Visscher, P.M. *et al.* (2008) Heritability in the genomics era--concepts and misconceptions, *Nat Rev Genet*, 9, 255-266.

Wellmann, R. and Bennewitz, J. (2012) Bayesian models with dominance effects for genomic evaluation of quantitative traits, *Genet Res*, 94, 21-37.

Yang, J. et al. (2011) GCTA: a tool for genome-wide complex trait analysis, *Am J Hum Genet*, 88(1):76-82

Zhou, X. and Matthew, S. (2012) Genome-wide efficient mixed-model analysis for association studies, *Nat Genet*, 44, 821-824.

Zhu, Z. *et al*. (2015) Dominance Genetic Variation Contributes Little to the Missing Heritability for Human Complex Traits, *The American Journal of Human Genetics*, 96, 377-385.