SI Appendix: Two-Way Mixed-Effects Methods for Joint Association Analysis Using Both Host and Pathogen Genomes

Supplementary Text

Supplementary Figures and Tables **S24**

www.pnas.org/cgi/doi/10.1073/pnas.1710980115

Supplementary Text

1 Two-Way Mixed-Effects Model for Association Analysis

An overview of the ATOMM method is provided in the paper. Here, we provide a detailed and self-contained description of our method, with some materials repeating from the paper if necessary. In Section 1.1, we overview the classical one-organism mixed-effects model and then propose an extension of the genetic relatedness matrix (GRM) that allows triallelic (or multiallelic) variants. In Section 1.2, we develop a two-way mixed-effects model for two-organism association analysis and describe the corresponding computational components. For ease of presentation, we focus mainly on the linear mixed model (LMM) for a quantitative trait, meaning the response is multivariate Gaussian. The extension of LMM to generalized linear mixed model (GLMM) with a non-Gaussian trait is described in Section 1.3, with a particular focus on a binomial-like trait.

1.1 One-Organism Linear Mixed Model

Suppose we observe both genotype and phenotype data from n individuals, where each individual represents an inbred host line or a haploid pathogen strain. Let Y_{ir} denote the quantitative trait measured on individual i and replicate r, where $i = 1, \ldots, n$ indexes the host line or pathogen strain and $r = 1, \ldots, k$ indexes replicates within each individual. In the genetic association study of a quantitative trait, Y_{ir} is typically modeled as conditionally multivariate normal:

$$Y_{ir} = X_{ir}\boldsymbol{\beta} + G_i^{\text{test}}\gamma + \eta_i + \varepsilon_{ir}, \quad \text{where} \quad \varepsilon_{ir} \sim \text{i.i.d. } N(0, \sigma_e^2), \tag{1}$$

where $X_{ir}\boldsymbol{\beta}$ represents the fixed effects of covariates, $G_i^{\text{test}}\gamma$ is the effect of the genetic variant currently being tested, η_i is the additive polygenic random effect of other variants in the genome not currently being tested (i.e., background variants), and ε_{ir} is i.i.d. Gaussian noise. The most common type of G_i^{test} is a single nucleotide polymorphism (SNP), in which G_i^{test} is encoded as either 0 or 1 for inbred lines $i = 1, \ldots, n$. We refer to G_i^{test} of this type as a "biallelic" variant. For a bacterial organism such as *X. arboricola*, its genome consists of core regions that are shared among all strains and dispensable regions that are present in only a subset of the sampled strains. In such a case, we also consider another type of G_i^{test} that takes three possible values, $G_i^{\text{test}} \in \{0, 1, D\}$, where D is an additional genotype status representing "**D**eletion". This essentially treats the "deletion" or "not of a site" as a 3rd type allele. Without loss of generality, we refer to G_i^{test} of this type as a "triallelic" variant. In equation (1), we could replace the term $G_i^{\text{test}}\gamma$ by $\gamma_D \mathbb{1}_{\{G_i^{\text{test}}=D\}} + \gamma_1 \mathbb{1}_{\{G_i^{\text{test}}=1\}}$ in that case, where $\mathbb{1}$ represents an indicator vector.

1.1.1 GRM Based on Biallelic Variants

In the case that all background variants are biallelic, the model for η_i, i.e., the additive polygenic effect of the variants in the genome for individual i, is written as

$$\eta_i = \sum_{l}^{m} \alpha_l \frac{G_{il} - f_l}{\sqrt{f_l(1-f_l)}}, \quad \text{for all } i = 1, \ldots, n, \tag{2}$$

where l indexes the variant in the genome, $G_{il} \in \{0, 1\}$ is the genotype of individual i at variant l, f_l is the allele frequency of variant l, $\frac{G_{il}-f_l}{\sqrt{f_l(1-f_l)}}$ is the standardized genotype of individual i at variant l, and the α_l's are i.i.d. random effects independent of $\boldsymbol{G}$ satisfying

$$\mathbb{E}(\alpha_l) = 0 \quad \text{and} \quad \text{Var}(\alpha_l) = \frac{1}{m}\sigma_a^2, \quad \text{for all } l = 1, \ldots, m. \tag{3}$$

Let $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)^T$, where n is the number of individuals in the study. Under regularity conditions sufficient for a Central Limit Theorem (CLT) as $m \to \infty$, the model equations (2) and (3) lead to the following asymptotic approximation for large m:

$$\boldsymbol{\eta}|\boldsymbol{G} \sim \mathcal{MVN}(\boldsymbol{0}, \sigma_a^2 \boldsymbol{K}), \tag{4}$$

where $\boldsymbol{K}$ is the GRM with (i, j)th entry

$$K(i, j) = \frac{1}{m} \sum_{l=1}^{m} \frac{(G_{il} - f_l)(G_{jl} - f_l)}{f_l(1 - f_l)}, \quad \text{for all } i, j = 1, \ldots, n. \tag{5}$$

In practice, the allele frequency, f_l, is unknown and we choose to use the sample average, $\hat{f}_l$, in place of f_l in (5). We focus on only *biallelic* variants (i.e., $G_{il} \in \{0, 1\}$) in the host *A. thaliana* genome, so the proposed GRM lends itself well to this context. Combining (1), (4) and (5) yields the one-organism LMM with two variance components:

$$\mathbb{E}(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{G}^{\text{test}}) = \boldsymbol{X}\boldsymbol{\beta} + (\boldsymbol{G}^{\text{test}} \otimes \boldsymbol{1}_k)\gamma \quad \text{and} \quad \text{Var}(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{G}^{\text{test}}) = \sigma_e^2 \boldsymbol{I} + \sigma_a^2 \left(\boldsymbol{K} \otimes \boldsymbol{1}_k \boldsymbol{1}_k^T\right), \tag{6}$$

where $\boldsymbol{Y}$, $\boldsymbol{X}$, $\boldsymbol{G}^{\text{test}}$ are vectorized versions of Y_{ir}, X_{ir} and G_i^{text}, respectively, $\boldsymbol{1}_k$ denotes a vector of length k with every element equal to 1, and $\boldsymbol{I}$ is an $n_{\text{obs}} \times n_{\text{obs}}$ identity matrix with $n_{\text{obs}} = nk$. Model (6) has four unknown parameters among which γ is the association parameter of interest and $\boldsymbol{\beta}$, σ_a^2, σ_e^2 are the nuisance parameters.

1.1.2 Extension to Multiallelic Variants

In the pathogen *X. arboricola* genome, a large number of the genetic variants we consider are effectively triallelic. This is because different strains of pathogen *X. arboricola* tend to have different genomic regions present. These regions form what is called the dispensable genome. The presence of the dispensable genome leads to many SNPs exhibiting three possible states among the sampled strains, i.e., $G_{il} \in \{0, 1, D\}$ for individual $i = 1, \ldots, n$. In such a case, the allele frequency, f_{il}, in the expression (5) is not well defined, so a more general model for η_i is needed.

Recall that the vector of trait phenotypes $\boldsymbol{Y} = (Y_1, \ldots, Y_{n_{\text{obs}}})^T$ can be modeled as conditionally multivariate normal:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{G}^{\text{test}}\gamma + \boldsymbol{Z}\boldsymbol{\eta} + \boldsymbol{\varepsilon}, \quad \text{where} \quad \boldsymbol{\varepsilon} \sim \mathcal{MVN}(\boldsymbol{0}, \sigma_e^2 \boldsymbol{I}), \tag{7}$$

where $\boldsymbol{Z}$ is the $n_{\text{obs}} \times n$ incidence matrix, $\boldsymbol{X}\boldsymbol{\beta}$ represents the fixed effects of covariates, $\boldsymbol{G}^{\text{test}}\gamma$ is the effect of the genetic variant currently being tested, $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)^T$ is the additive polygenic random effect of other variants in the genome not currently being tested (i.e., background variants). In the case when all background variants are *triallelic*, we propose to decompose the additive polygenic random effect, η_i, of individual i into two orthogonal parts,

$$\eta_i = \eta_{iS} + \eta_{iD}, \quad \text{for all } i = 1, \ldots, n, \tag{8}$$

where η_{iS} is the random effect due to SNP alleles, η_{iD} is the random effect due to deletion or not of sites, where both η_{iS} and η_{iD} have similar structure as (2) and (3). Specifically, let $l = 1, \ldots, m$ index the background variant, and let $f_{ls} \in (0, 1)$ be the frequency of state $s \in \{0, 1, D\}$ at variant l, where $f_{l1} + f_{l0} + f_{lD} = 1$. We write

$$\eta_{iD} = \sum_{l=1}^{m} \alpha_{lD} \underbrace{\frac{\mathbb{1}_{\{G_{il}=D\}} - f_{lD}}{\sqrt{f_{lD}(1 - f_{lD})}}}_{\stackrel{\text{def}}{=} A_{il}}, \tag{9}$$

and

$$\eta_{iS} = \sum_{l=1}^{m} \alpha_{lS} \underbrace{\left(\frac{1}{\sqrt{1 - f_{lD}}} \frac{\mathbb{1}_{\{G_{il}=1\}} - p_l}{\sqrt{p_l(1 - p_l)}} \right) \mathbb{1}_{\{G_{il} \neq D\}}}_{\stackrel{\text{def}}{=} B_{il}}, \quad \text{with} \quad p_l \stackrel{\text{def}}{=} \mathbb{P}\left(\mathbb{1}_{\{G_{il}=1\}} | \mathbb{1}_{\{G_{il} \neq D\}} \right) = \frac{f_{l1}}{1 - f_{lD}}, \tag{10}$$

for all $i = 1, \ldots, n$, where $\mathbb{1}_{\{G_{il}=s\}}$ denotes the indicator function for $s \in \{0, 1, D\}$, A_{il}, B_{il} are the standardized genotypes of individual i at variant l, and α_{lS}, α_{lD} are both i.i.d. random effects (to be specified later) for variant $l = 1, \ldots m$. We note that by construction, $n^{-1/2}(A_{1l}, \ldots, A_{nl})^T$ and $n^{-1/2}(B_{1l}, \ldots, B_{nl})^T$ are orthonormal vectors in $\mathbb{R}^n$ for all $l = 1, \ldots, m$, if we plug in $\hat{f}_{lD} = n^{-1} \sum_{i=1}^{n} \mathbb{1}_{\{G_{il}=D\}}$, $\hat{f}_{l1} = n^{-1} \sum_{i=1}^{n} \mathbb{1}_{\{G_{il}=1\}}$, and $\hat{p}_l = \frac{\hat{f}_{l1}}{1 - \hat{f}_{lD}}$ for f_{lD}, f_{l1} and p_l, respectively.

Now similarly as in (3), we make the modeling assumptions that $\alpha_{1S}, \ldots, \alpha_{mS}$ are i.i.d. random effects with

$$\mathbb{E}(\alpha_{lS}) = 0, \ \text{Var}(\alpha_{lS}) = \frac{1}{m}\sigma_S^2, \quad \text{for } l = 1, \ldots, m,$$

$\alpha_{1D}, \ldots, \alpha_{mD}$ are i.i.d. random effects with

$$\mathbb{E}(\alpha_{lD}) = 0, \ \text{Var}(\alpha_{lD}) = \frac{1}{m}\sigma_D^2, \quad \text{for } l = 1, \ldots, m,$$

and $\alpha_{1S}, \ldots, \alpha_{mS}, \alpha_{1D}, \ldots, \alpha_{mD}$ are mutually independent and independent of $\boldsymbol{G}$. If we let $\boldsymbol{\eta}_S = (\eta_{1S}, \ldots, \eta_{nS})^T$ and $\boldsymbol{\eta}_D = (\eta_{1D}, \ldots, \eta_{nD})^T$, then under regularity conditions sufficient for a CLT, we obtain the following asymptotic approximation for large m:

$$\boldsymbol{\eta}_S | \boldsymbol{G} \sim \mathcal{MVN}(\boldsymbol{0}, \sigma_S^2 \boldsymbol{M}_S), \quad \boldsymbol{\eta}_D | \boldsymbol{G} \sim \mathcal{MVN}(\boldsymbol{0}, \sigma_D^2 \boldsymbol{M}_D), \quad \text{and} \quad \boldsymbol{\eta}_S \perp \boldsymbol{\eta}_D | \boldsymbol{G}, \tag{11}$$

where $\boldsymbol{M}_D$ and $\boldsymbol{M}_S$ are variations on the biallelic GRM (5) that are constructed from (9) and (10).

A parallel explanation. We could write (7) as

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{G}^{\text{test}}\gamma + \boldsymbol{Z}\underbrace{\left(\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)\boldsymbol{G}_D\boldsymbol{W}_D}_{\stackrel{\text{def}}{=}\boldsymbol{A}}\boldsymbol{\alpha}_D + \boldsymbol{Z}\underbrace{\left[\left(\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)\boldsymbol{G}_S\boldsymbol{W}_S\right] \odot \left[\left(\boldsymbol{I} - \boldsymbol{G}_D\right)\boldsymbol{V}_D\right]}_{\stackrel{\text{def}}{=}\boldsymbol{B}}\boldsymbol{\alpha}_S + \boldsymbol{\varepsilon}, \quad (12)$$

where $\odot$ denotes Hadamard product, $\boldsymbol{Z}$ is the $n_{\text{obs}} \times n$ incidence matrix, $\boldsymbol{G}_D$ is the $n \times m$ genotype incidence matrix with (i,l)th entry equal to $\mathbb{1}_{\{G_{il}=D\}}$, $\boldsymbol{G}_S$ is the $n \times m$ genotype incidence matrix with (i,l)th entry equal to $\mathbb{1}_{\{G_{il}=1\}}$, $\boldsymbol{W}_D$ is the $m \times m$ diagonal with (l,l)th entry equal to $\frac{1}{\sqrt{f_{lD}(1-f_{lD})}}$, $\boldsymbol{W}_S$ is an $m \times m$ diagonal matrix with (l,l)th entry equal to $\frac{1}{\sqrt{p_l(1-p_l)}}$, $\boldsymbol{V}_D$ is a length-m vector with lth entry equal to $\frac{1}{\sqrt{1-f_{lD}}}$, $\boldsymbol{\alpha}_D = (\alpha_{1D}, \ldots, \alpha_{mD})^T$, $\boldsymbol{\alpha}_S = (\alpha_{1S}, \ldots, \alpha_{mS})^T$, and we make modeling assumptions that $\boldsymbol{\varepsilon}$, $\boldsymbol{\alpha}_D$ and $\boldsymbol{\alpha}_S$ are independent,

$$\mathbb{E}(\boldsymbol{\alpha}_D) = \mathbf{0}, \quad \text{Var}(\boldsymbol{\alpha}_D) = \frac{\sigma_D^2}{m}\boldsymbol{I}, \quad \mathbb{E}(\boldsymbol{\alpha}_S) = \mathbf{0}, \quad \text{Var}(\boldsymbol{\alpha}_S) = \frac{\sigma_S^2}{m}\boldsymbol{I}, \quad \mathbf{E}(\boldsymbol{\varepsilon}) = \mathbf{0}, \quad \text{Var}(\boldsymbol{\varepsilon}) = \sigma_e^2\boldsymbol{I}.$$

Then the model equation (12) implies

$$\text{Var}(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{G}^{\text{test}}, \boldsymbol{Z}, \boldsymbol{G}_D, \boldsymbol{G}_S) = \sigma_D^2\boldsymbol{M}_D + \sigma_S^2\boldsymbol{M}_S + \sigma_e^2\boldsymbol{I},$$

where

$$\boldsymbol{M}_D = \frac{1}{m}\boldsymbol{Z}(\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T)\boldsymbol{G}_D\boldsymbol{W}_D^2\boldsymbol{G}_D^T(\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T)\boldsymbol{Z}^T = \frac{1}{m}\boldsymbol{Z}\boldsymbol{A}\boldsymbol{A}^T\boldsymbol{Z}^T, \quad \text{and} \quad \boldsymbol{M}_S = \frac{1}{m}\boldsymbol{Z}\boldsymbol{B}\boldsymbol{B}^T\boldsymbol{Z}^T. \quad \square$$

For parsimony, we take $\sigma_S^2 = \sigma_D^2$ and call it $\sigma_a^2/2$. With some bookkeeping, the model equations (8), (9), (10) and (11) reduce to

$$\boldsymbol{\eta} \sim \mathcal{MVN}(\mathbf{0}, \sigma_a^2\boldsymbol{K}), \quad (13)$$

where $\boldsymbol{K} = (\boldsymbol{M}_S + \boldsymbol{M}_D)/2$ is the GRM based on triallelic variants, with (i,j)th entry

$$K(i,j) = \frac{1}{2m}\sum_{l=1}^{m}\left(-\mathbb{1}_{\{G_{il}\neq G_{jl}\}} + \sum_{s\in\{0,1,D\}}\frac{1-f_{ls}}{f_{ls}}\mathbb{1}_{\{G_{il}=G_{jl}=s\}}\right), \quad \text{for all } i,j = 1,\ldots,n. \quad (14)$$

The factor of 2 in (14) is to ensure that $\mathbb{E}(K_{ii}) = 1$ under the assumption that G_{il} follows a 3-class categorical distribution. Finally, in the general case when the background variants consist of both biallelic and triallelic variants, we construct the empirical GRM using the weighted average of (5) and (14), where the weight is proportional to the number of corresponding variants, and the sample frequencies are used in place of the true allele frequencies to construct $\boldsymbol{K}$.

More generally, if all m genetic variants had δ alleles, $\delta \geq 2$, then we could use equation (13) with $\boldsymbol{K}$ having (i,j)th entry

$$K(i,j) = \frac{1}{(\delta-1)m}\sum_{l=1}^{m}\left(-\mathbb{1}_{\{G_{il}\neq G_{jl}\}} + \sum_{s=1}^{\delta}\frac{1-f_{ls}}{f_{ls}}\mathbb{1}_{\{G_{il}=G_{jl}=s\}}\right). \quad (15)$$

Note that in the case $\delta = 2$, equation (15) reduces to the commonly-used biallelic GRM given in (5).

From the viewpoint of modeling genotypes, a connection between equation (14) and kinship estimation is

given in the following theorem:

Theorem 1.1. *Fix a genetic variant $l \in \{1, \ldots, m\}$. Assume that for individuals $i = 1, \ldots, n$, G_{il} follows a 3-class categorical distribution with occurrence frequencies $0 < f_{ls} < 1$ for $s \in \{0, 1, D\}$, where $f_{l0}+f_{l1}+f_{lD} = 1$. Let $\phi(i, j)$ denote the kinship coefficient between individuals i and j. Consider a class of linear estimators of the form*

$$\hat{\phi}(i,j) = \beta_{l4}\mathbb{1}_{\{G_{il}\neq G_{jl}\}} + \sum_{s\in\{0,1,D\}} \beta_{ls}\mathbb{1}_{\{G_{il}=G_{jl}=s\}}, \tag{16}$$

where $\beta_{l4} \in \mathbb{R}$ and $\beta_{ls} \in \mathbb{R}$ for $s \in \{0, 1, D\}$. If $\phi(i, j) = \delta > 0$, then as $\delta \to 0$, the minimum-variance unbiased estimator (MVUE) in the class (16) is

$$\hat{\phi}^*(i,j) \to -\frac{1}{2}\mathbb{1}_{\{G_{il}\neq G_{jl}\}} + \sum_{s\in\{0,1,D\}} \frac{1-f_{ls}}{2f_{ls}}\mathbb{1}_{\{G_{il}=G_{jl}=s\}}.$$

Proof. Consider a pair of individuals, i, j, and their alleles at variant l, G_{il}, G_{jl}. For simplicity, we drop the index l in the proof. By definition, the probability for the event $\{G_i$ and G_j are identical by descent$\}$ is $\phi(i, j)$. By (16), we have

$$\begin{aligned}
\mathbb{E}[\hat{\phi}(i,j)] &= \beta_4\mathbb{E}\mathbb{1}_{\{G_i\neq G_j\}} + \sum_{s\in\{0,1,D\}} \beta_s\mathbb{E}\mathbb{1}_{\{G_i=G_j=s\}} \\
&= \beta_4\left[1-\phi(i,j)\right]\left(1-f_0^2-f_1^2-f_D^2\right) + \sum_{s\in\{0,1,D\}} \beta_s\left\{\phi(i,j)f_s + \left[1-\phi(i,j)\right]f_s^2\right\} \\
&= \phi(i,j)\underbrace{\left[\sum_{s\in\{0,1,D\}} \beta_s f_s(1-f_s) - \beta_4\left(1-\sum_{s\in\{0,1,D\}} f_s^2\right)\right]}_{\text{Part I}} + \underbrace{\sum_{s\in\{0,1,D\}} \beta_s f_s^2 + \beta_4\left(1-\sum_{s\in\{0,1,D\}} f_s^2\right)}_{\text{Part II}}.
\end{aligned}$$

Since we are interested in the unbiased estimator, we set $\mathbb{E}[\hat{\phi}(i, j)] = \phi(i, j) \neq 0$. This implies Part I $= 1$ and Part II $= 0$. So, $(\beta_0, \beta_1, \beta_D, \beta_4)$ satisfies the following equations

$$\begin{cases} \sum_{s\in\{0,1,D\}} \beta_s f_s(1-f_s) - \beta_4\left(1-\sum_{s\in\{0,1,D\}} f_s^2\right) = 1, \\ \sum_{s\in\{0,1,D\}} \beta_s f_s^2 + \beta_4\left(1-\sum_{s\in\{0,1,D\}} f_s^2\right) = 0. \end{cases} \tag{17}$$

Simplifying (17) gives

$$\begin{cases} \beta_0 f_0 + \beta_1 f_1 + \beta_D f_D = 1, \\ \beta_4 = -\dfrac{\beta_0 f_0^2 + \beta_1 f_1^2 + \beta_D f_D^2}{1-f_0^2-f_1^2-f_D^2}. \end{cases} \tag{18}$$

For the variance of $\hat{\phi}(i, j)$, we have

$$\begin{aligned}
\mathrm{Var}[\hat{\phi}(i,j)] &= \mathbb{E}[\hat{\phi}^2(i,j)] - \mathbb{E}[\hat{\phi}(i,j)]^2 \\
&= \sum_{s\in\{0,1,D\}} \beta_s^2\mathbb{E}\mathbb{1}_{\{G_i=G_j=s\}} + \beta_4^2\mathbb{E}\mathbb{1}_{\{G_i\neq G_j\}} - \phi^2(i,j) \\
&= \sum_{s\in\{0,1,D\}} \beta_s^2\left\{\phi(i,j)f_s + \left[1-\phi(i,j)f_s^2\right]\right\} + \beta_4^2\left[1-\phi(i,j)\right]\left(1-f_0^2-f_1^2-f_D^2\right) - \phi^2(i,j)
\end{aligned} \tag{19}$$

Plugging (18) into (19), we obtain

$$\mathrm{Var}[\hat{\phi}(i,j)] = \phi(i,j) \sum_{s\in\{0,1,D\}} \beta_s^2 f_s + [1-\phi(i,j)] \left[\sum_{s\in\{0,1,D\}} \beta_s^2 f_s^2 + \frac{\left(\beta_0 f_0^2 + \beta_1 f_1^2 + \beta_D f_D^2\right)^2}{1 - f_0^2 - f_1^2 - f_D^2} \right] - \phi^2(i,j). \quad (20)$$

In order to find $\hat{\phi}^*(i,j)$, the MVUE of the form (16), we seek to solve the following optimization problem:

$$\begin{aligned} \min_{\beta_0,\beta_1,\beta_D\in\mathbb{R}} \quad & \mathrm{Var}[\hat{\phi}(i,j)] \\ \text{subject to} \quad & \beta_0 f_0 + \beta_1 f_1 + \beta_D f_D = 1, \end{aligned} \quad (21)$$

where the second line follows from (18). We note that there does not exist a single $\hat{\phi}^*(i,j)$ that uniformly solves (21) for all $\phi(i,j) \in [0,1]$. Instead, we consider the following boundary case when $\phi(i,j) = \delta \to 0$. Then (20) simplifies to

$$\mathrm{Var}[\hat{\phi}(i,j)] = \left(\sum_{s\in\{0,1,D\}} \beta_s^2 f_s^2 \right) + \left(1 - \sum_{s\in\{0,1,D\}} f_s^2 \right)^{-1} \left(\sum_{s\in\{0,1,D\}} \beta_s f_s^2 \right)^2 + O(\delta). \quad (22)$$

We apply the Langrange multiplier method to solve the constrained optimization (21). Specifically, define

$$H(\beta_0,\beta_1,\beta_D,\lambda) = \left(1 - \sum_{s\in\{0,1,D\}} f_s^2 \right) \left(\sum_{s\in\{0,1,D\}} \beta_s^2 f_s^2 \right) + \left(\sum_{s\in\{0,1,D\}} \beta_s f_s^2 \right)^2 - \lambda \left(\sum_{s\in\{0,1,D\}} \beta_s f_s - 1 \right). \quad (23)$$

Setting the partial derivatives to zero gives

$$\begin{cases} 0 = \frac{\partial H}{\partial \beta_s} = 2C f_s^2 \beta_s + 2E f_s^2 - \lambda f_s, \quad \text{for } s \in \{0,1,D\}, \\ 0 = \frac{\partial H}{\partial \lambda} = \sum_{s\in\{0,1,D\}} \beta_s f_s - 1, \end{cases} \quad \text{where} \quad \begin{cases} C \overset{\text{def}}{=} 1 - \sum_{s\in\{0,1,D\}} f_s^2, \\ E = E(\beta_0,\beta_1,\beta_D) \overset{\text{def}}{=} \sum_{s\in\{0,1,D\}} \beta_s f_s^2. \end{cases} \quad (24)$$

Since $f_s \neq 0$, the first equation in (24) implies

$$2C\beta_s f_s + 2E f_s = \lambda, \quad \text{for } s \in \{0,1,D\}. \quad (25)$$

Combining this with the second equation in (24) and the definitions of C and E, we obtain

$$\begin{cases} \sum_{s\in\{0,1,D\}} (2C\beta_s f_s + 2E f_s) = \sum_{s\in\{0,1,D\}} \lambda, \\ \sum_{s\in\{0,1,D\}} \left(2C\beta_s f_s^2 + 2E f_s^2\right) = \sum_{s\in\{0,1,D\}} \lambda f_s, \end{cases} \Rightarrow \begin{cases} 2C + 2E = 3\lambda, \\ 2CE + 2E(1-C) = \lambda, \end{cases} \Rightarrow \begin{cases} C = \lambda, \\ E = \frac{\lambda}{2}. \end{cases} \quad (26)$$

Now pluggging (26) back into (25) yields

$$\beta_s = \frac{\lambda - 2E f_s}{2C f_s} = \frac{1 - f_s}{2 f_s}, \quad \text{for } s \in \{0,1,D\}.$$

Note that the constrained optimization (22) is equivalent to the unconstrained optimization (23) up to a

small factor $O(\delta) = o(1)$. Therefore, the minimizer of (22) under the constraint (21) is

$$\beta_s^* = \frac{1 - f_s}{2f_s} + o(1), \quad \text{for } s \in \{0, 1, D\}.$$

By (18) and (16), this implies that the MVUE in this case is

$$\hat{\phi}^*(i,j) = -\frac{1}{2}\mathbb{1}_{\{G_{il} \neq G_{jl}\}} + \sum_{s \in \{0,1,D\}} \frac{1 - f_{ls}}{2f_{ls}} \mathbb{1}_{\{G_{il} = G_{jl} = s\}} + o(1). \qquad \square$$

Theorem 1.1 elucidates that the $K_{i,j}$ we propose in (14) is an unbiased estimate of $\phi(i,j)$, and would have close to the minimum variance if the true kinship coefficient $\phi(i,j)$ were close to 0. This property seems desirable because the levels of relatedness are often low in natural populations, and the presence of non-zero relatedness increases only the variance of $\hat{\phi}^*(i,j)$ but not its bias (in the case when (f_0, f_s, f_D) is known).

1.2 Gaussian ATOMM: Two-Organism Linear Mixed Model

We now have all the ingredients necessary to describe the conditionally Gaussian version of ATOMM. Suppose we observe both genotype and phenotype data from n host-pathogen pairs. For simplicity, we ignore the replicates for the moment. Let Y_{ij} denote the trait value measured on host-pathogen pair (i,j), where $i = 1, \cdots, n_h$ indexes the host inbred line, and $j = 1, \cdots, n_p$ indexes the pathogen strain, with $n_h n_p = n$. We propose to model Y_{ij} as

$$\underbrace{Y_{ij}}_{\text{Response}} = \underbrace{X_{ij}\boldsymbol{\beta}}_{\text{Covariates}} + \underbrace{G_i^{h,\text{test}}\gamma_1 + G_j^{p,\text{test}}\gamma_2 + G_i^{h,\text{test}}G_j^{p,\text{test}}\gamma_3}_{\text{Fixed Effects of Interest}} + \underbrace{\eta_i^h + \eta_j^p + \eta_{ij}^{hp}}_{\text{Random Effects}} + \underbrace{\varepsilon_{ij}}_{\text{i.i.d. Noise}}, \tag{27}$$

where $X_{ij}\boldsymbol{\beta}$ represents covariate effects, $G_i^{h,\text{test}}\gamma_1$ is the effect of the host genetic variant being tested, $G_j^{p,\text{test}}\gamma_2$ is the effect of the pathogen genetic variant being tested, $G_i^{h,\text{test}}G_j^{p,\text{test}}\gamma_3$ is the interaction effect between the tested host variant and the tested pathogen variant, η_i^h, η_i^p and η_{ij}^{hp} are additive polygenic random effects (to be specified later) of other variants not currently being tested, and ε_{ij} is assumed i.i.d. $N(0, \sigma_e^2)$. Motivated by the features of the *A. thaliana* and *X. arboricola* genomes, we consider $G_i^{h,\text{test}} \in \{0,1\}$, i.e., each host variant has two possible genotype states, whereas either $G_j^{p,\text{test}} \in \{0,1\}$ or $G_j^{p,\text{test}} \in \{0,1,D\}$, depending on whether the pathogen genetic variant under consideration is biallelic or triallelic. In the latter case, we replace the term $G_j^{p,\text{test}}\gamma_2$ by $\gamma_{2,D}\mathbb{1}_{\{G_j^{p,\text{test}}=D\}} + \gamma_{2,1}\mathbb{1}_{\{G_j^{p,\text{test}}=1\}}$, and we replace the term $G_i^{h,\text{test}}G_j^{p,\text{test}}\gamma_3$ by $\gamma_{3,D}G_i^{h,\text{test}}\mathbb{1}_{\{G_j^{p,\text{test}}=D\}} + \gamma_{3,1}G_i^{h,\text{test}}\mathbb{1}_{\{G_j^{p,\text{test}}=1\}}$ in equation (27). Here the encoding of the genotypic values is appropriate for genetic variants of inbred *A. thaliana* lines or haploid *X. arboricola* strains. With little modification, our method can easily extend to other encoding of genotypic values, e.g., for diploid individuals, depending on the specific host/pathogen organisms in the study.

To model the random effects, let $\boldsymbol{\eta}^h = (\eta_1^h, \ldots, \eta_{n_h}^h)^T$ represent the host additive polygenic random effects, $\boldsymbol{\eta}^p = (\eta_1^p, \ldots, \eta_{n_p}^p)^T$ be the pathogen additive polygenic random effects, and $\boldsymbol{\eta}^{hp} = (\eta_{11}^{hp}, \ldots, \eta_{1n_p}^{hp}, \ldots, \eta_{n_h1}^{hp}, \ldots, \eta_{n_hn_p}^{hp})^T$ be the host-pathogen additive-by-additive interaction random effects. Following the same

lines as in Sections 1.1.1 and 1.1.2, we propose the model:

$$\boldsymbol{\eta}^h \sim \mathcal{MVN}(\mathbf{0}, \sigma_h^2 \boldsymbol{K}_h), \quad \boldsymbol{\eta}^p \sim \mathcal{MVN}(\mathbf{0}, \sigma_p^2 \boldsymbol{K}_p), \quad \text{and} \quad \boldsymbol{\eta}^{hp} \sim \mathcal{MVN}(\mathbf{0}, \sigma_{hp}^2 \boldsymbol{K}_{hp}), \tag{28}$$

where $\boldsymbol{K}_h$ is the host GRM described in Section 1.1.1, $\boldsymbol{K}_p$ is the pathogen GRM described in Section 1.1.2, and $\boldsymbol{K}_{hp}$ is the covariance matrix for the host-pathogen interaction effects. We propose to model $\boldsymbol{K}_{hp}$ as

$$\boldsymbol{K}_{hp} = \boldsymbol{K}_h \otimes \boldsymbol{K}_p, \tag{29}$$

where $\otimes$ denotes the Kronecker product. This choice of $\boldsymbol{K}^{hp}$ can be derived as above using Fisher's infinitesimal approach. Specifically, let η_{ij}^{hp} denote the interaction random effect for host i and pathogen j. In the simplest case where host and pathogen variants both are biallelic, we propose

$$\eta_{ij}^{hp} = \sum_{k=1}^{m_h} \sum_{l=1}^{m_p} \beta_{lk} \frac{(G_{ki}^h - f_k^h)}{\sqrt{f_k^h(1-f_k^h)}} \frac{(G_{lj}^p - f_l^p)}{\sqrt{f_l^p(1-f_l^p)}}, \quad \text{for all } i = 1, \ldots, n_h, \quad j = 1, \ldots, n_p,$$

where β_{lk} are i.i.d. random effects with $\mathbb{E}(\beta_{lk}) = 0$, $\mathrm{Var}(\beta_{lk}) = m_h^{-1} m_p^{-1} \sigma_{hp}^2$. If we let $\boldsymbol{\eta}^{hp} = (\eta_{11}^{hp}, \ldots, \eta_{1n_p}^{hp}, \ldots, \eta_{n_h 1}^{hp}, \ldots, \eta_{n_h n_p}^{hp})^T$, then under regularity conditions sufficient for a CLT, we obtain the asymptotic approximation $\boldsymbol{\eta} \sim \mathcal{MVN}(\mathbf{0}, \sigma_{hp}^2 \boldsymbol{K}_{hp})$ for large $m_h \cdot m_p$, where $\boldsymbol{K}_{hp} = \boldsymbol{K}_h \otimes \boldsymbol{K}_p$. Similarly, we also use $\eta \sim \mathcal{MVN}(\mathbf{0}, \sigma_{hp}^2 \boldsymbol{K}_{hp})$ with $\boldsymbol{K}_{hp} = \boldsymbol{K}_h \otimes \boldsymbol{K}_p$ when the pathogen genome containing both biallelic and triallelic variants, where in that case, we take $\boldsymbol{K}_p$ to be the weighted average of (5) and (14), with the weight proportional to the number of corresponding variants.

By (27), (28) and (29), the vectorized version of the full model, ignoring replicates, would be

$$\begin{aligned} \boldsymbol{Y} | \boldsymbol{X}, \boldsymbol{G}^{h,\text{test}}, \boldsymbol{G}^{h,\text{test}} &\sim \mathcal{MVN}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad \text{where} \\ \boldsymbol{\mu} &= \boldsymbol{X}\boldsymbol{\beta} + \left(\boldsymbol{G}^{h,\text{test}} \otimes \mathbf{1}_{n_p}\right) \gamma_1 + \left(\mathbf{1}_{n_p} \otimes \boldsymbol{G}^{p,\text{test}}\right) \gamma_2 + \left(\boldsymbol{G}^{h,\text{test}} \otimes \boldsymbol{G}^{p,\text{test}}\right) \gamma_3, \\ \boldsymbol{\Sigma} &= \sigma_h^2 [\boldsymbol{K}_h \otimes (\mathbf{1}_{n_p} \mathbf{1}_{n_p}^T)] + \sigma_p^2 [(\mathbf{1}_{n_p} \mathbf{1}_{n_p}^T) \otimes \boldsymbol{K}_p] + \sigma_{hp}^2 (\boldsymbol{K}_h \otimes \boldsymbol{K}_p) + \sigma_2^2 \boldsymbol{I}. \end{aligned} \tag{30}$$

There are four variance components in the model, with the covariance matrices, $\boldsymbol{K}_h$, $\boldsymbol{K}_p$, $\boldsymbol{K}_{hp}$, reflecting the polygenic effects of host, pathogen, and their interactions, respectively. We use sample allele frequencies in place of f_l and $f_{ls}, s \in \{0, 1, D\}$ to estimate $\boldsymbol{K}_h$, $\boldsymbol{K}_p$ and $\boldsymbol{K}_{hp}$.

1.3 Binomial-like ATOMM: Extension to non-Gaussian Phenotype

In situations when the phenotype of interest is non-Gaussian, one could still apply the Gaussian ATOMM though one might expect the mean to be related to the variance. Alternatively, one can extend (30) by taking a quasi-likelihood approach to circumvent the necessity of specifying a full probability model. Specifically, instead of specifying the full distribution as in (27), we consider a semi-parametric model for $\boldsymbol{Y}$ by specifying only the first two conditional moments. For example, our motivating *A. thaliana–X. arboricola* dataset has a binomial-like response, $y_{ij} \in \{0, 1, 2, \ldots, k\}$ (here, $k = 4$), so one could consider the following model for the mean

$$\begin{aligned} &\mathrm{E}(\boldsymbol{Y} | \boldsymbol{X}, \boldsymbol{G}^{h,\text{test}}, \boldsymbol{G}^{p,\text{test}}) = k\boldsymbol{\mu} \quad \text{where} \\ &\text{logit}\,(\boldsymbol{\mu}) = \boldsymbol{X}\boldsymbol{\beta} + (\boldsymbol{G}^{h,\text{test}} \otimes \mathbf{1}_{n_p})\gamma_1 + (\mathbf{1}_{n_a} \otimes \boldsymbol{G}^{p,\text{test}})\gamma_2 + (\boldsymbol{G}^{h,\text{test}} \otimes \boldsymbol{G}^{p,\text{test}})\gamma_3, \end{aligned} \tag{31}$$

and for the variance

$$\begin{aligned}&\mathrm{Var}(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{G}^{h,\mathrm{test}},\boldsymbol{G}^{p,\mathrm{test}})=\sigma_t^2\boldsymbol{M}\boldsymbol{\Sigma}\boldsymbol{M},\quad \text{where}\\&\boldsymbol{\Sigma}=\xi_1\boldsymbol{K}_h\otimes(\mathbf{1}_{n_p}\mathbf{1}_{n_p}^T)+\xi_2\left(\mathbf{1}_{n_h}\mathbf{1}_{n_h}^T\right)\otimes\boldsymbol{K}_p+\xi_3\boldsymbol{K}_{hp}+(1-\xi_1-\xi_2-\xi_3)\boldsymbol{I},\end{aligned}\tag{32}$$

where $\boldsymbol{M}$ is a diagonal matrix with ith diagonal element equal to $\sqrt{k\mu_i(1-\mu_i)}$; (ξ_1,ξ_2,ξ_3) parameterizes the proportion of total variance explained by each component and $0\le\xi_i\le 1,\ i=1,2,3,\ 0\le\xi_1+\xi_2+\xi_3\le 1$; $\boldsymbol{\Sigma}$ is pre- and post-multiplied by the diagonal matrix $\boldsymbol{M}$ so that the conditional variance of Y_i is $k\mu_i(1-\mu_i)\sigma_t^2$; and σ_t^2 is an additional unknown parameter for dispersion.

The model specified in (31) and (32) is a natural generalization of the two-way mixed-effects LMM (30) to the two-way mixed-effects GLMM. Using the quasi-likelihood framework, model (32) can be further generalized to allow other types of trait, such as a binary or Poisson-like trait. In general, one can choose a suitable link function and a diagonal matrix $\boldsymbol{M}$ of the form $\boldsymbol{M}=\mathrm{diag}\{\sqrt{V(\mu_1)},\cdots,\sqrt{V(\mu_n)}\}$ where the variance function $V(\cdot)$ is chosen based on the conditional variance of Y_i in the corresponding exponential family.

In situations in which the phenotype is non-Gaussian, GLMM seems to better reflect the nature of the phenotype distribution. In the *A. thaliana–X. arboricola* analysis, we found that the variance-mean relationship for QDR resembles the binomial variance function with dispersion allowed, $V(\mu)=4\mu(1-\mu)\sigma^2$. However, a Gaussian linear approximation usually works well in practice [2, 3, 13]. In fact, the linear model on a binary or binomial phenotype finds broad use in GWAS and has been shown to be rather robust to model misspecification. On the other hand, a careful use of a logistic link function and variance function could potentially increase power, especially when covariate effects are large [12]. In Section 3.5, we compare these two approaches by assessing the association results in the *A. thaliana–X. arboricola* dataset.

2 Parameter Estimation by Solution of Estimating Equations

Following a similar approach to that of Jiang et al. [12] and Zhong et al. [25], parameter estimation for the quasi-likelihood model (30) can be performed under either the full model given in (30) or under the null hypothesis

$$\mathcal{H}_0:\gamma_1=\gamma_2=\gamma_3=0.\tag{33}$$

In the context of the score tests for association that we perform, it is the parameter values estimated under the null that are needed. We therefore present a scheme for parameter estimation under the null for the binomial-like ATOMM (Section 1.3). From that, the procedure for the Gaussian ATOMM (Section 1.2) follows naturally, because the quasi-likelihood model for Gaussian ATOMM has similar expressions to those in (31) and (32), except that we replace the logistic link by the identity function and set $\boldsymbol{M}=\boldsymbol{I}$.

2.1 System of Estimating Equations

In the null model, the parameters to be estimated are $\boldsymbol{\beta}$, σ_t^2, ξ_1, ξ_2 and ξ_3. For the binomial-like model, we have specified only first and second moments, rather than specifying the entire distribution. We therefore use an estimating equation approach [12, 25] for parameter estimation under the null. In this approach, the fixed effects are estimated using quasi-likelihood, with additional estimating equations constructed for the variance

components, based on setting observed values of certain quadratic forms equal to their expectations. The resulting estimating equations for the fixed effects $\boldsymbol{\beta}$ and variance components σ_t^2, ξ_1, ξ_2, ξ_3 can be written as

$$\begin{cases} \boldsymbol{X}^T\boldsymbol{M}\boldsymbol{\Sigma}^{-1}\boldsymbol{M}^{-1}(\boldsymbol{Y}-k\boldsymbol{\mu})=0, \\ \sigma_t^2=n^{-1}(\boldsymbol{Y}-k\boldsymbol{\mu})^T\boldsymbol{M}^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{M}^{-1}(\boldsymbol{Y}-k\boldsymbol{\mu}), \\ \sigma_t^{-2}(\boldsymbol{Y}-k\boldsymbol{\mu})^T\boldsymbol{M}^{-1}\boldsymbol{\Sigma}^{-1}[\boldsymbol{K}_h\otimes(\mathbf{1}_{n_p}\mathbf{1}_{n_p}^T)-\boldsymbol{I}]\boldsymbol{\Sigma}^{-1}\boldsymbol{M}^{-1}(\boldsymbol{Y}-k\boldsymbol{\mu})=\mathrm{tr}\left\{\boldsymbol{\Sigma}^{-1}[\boldsymbol{K}_h\otimes(\mathbf{1}_{n_p}\mathbf{1}_{n_p}^T)-\boldsymbol{I}]\right\}, \\ \sigma_t^{-2}(\boldsymbol{Y}-k\boldsymbol{\mu})^T\boldsymbol{M}^{-1}\boldsymbol{\Sigma}^{-1}\left[(\mathbf{1}_{n_h}\mathbf{1}_{n_h}^T)\otimes\boldsymbol{K}_p-\boldsymbol{I}\right]\boldsymbol{\Sigma}^{-1}\boldsymbol{M}^{-1}(\boldsymbol{Y}-k\boldsymbol{\mu})=\mathrm{tr}\left\{\boldsymbol{\Sigma}^{-1}[(\mathbf{1}_{n_h}\mathbf{1}_{n_h}^T)\otimes\boldsymbol{K}_p-\boldsymbol{I}]\right\}, \\ \sigma_t^{-2}(\boldsymbol{Y}-k\boldsymbol{\mu})^T\boldsymbol{M}^{-1}\boldsymbol{\Sigma}^{-1}(\boldsymbol{K}_{hp}-\boldsymbol{I})\boldsymbol{\Sigma}^{-1}\boldsymbol{M}^{-1}(\boldsymbol{Y}-k\boldsymbol{\mu})=\mathrm{tr}\left[\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{K}_{hp}-\boldsymbol{I}\right)\right]. \end{cases} \tag{34}$$

We solve the above equations and take the solutions, $\hat{\boldsymbol{\beta}}_0$, $\hat{\sigma}_{t,0}$, $\hat{\xi}_{1,0}$,$\hat{\xi}_{2,0}$ and $\hat{\xi}_{3,0}$, as the null estimates. Here, the subscript "0" means that estimation is under the null (33). Because the system (34) involves non-linear equations, a closed form solution is not known in general. We take a numerical search approach [12, 25] to solve for $\hat{\boldsymbol{\beta}}_0$, $\hat{\sigma}_{t,0}$, $\hat{\xi}_{1,0}$,$\hat{\xi}_{2,0}$ and $\hat{\xi}_{3,0}$.

2.2 Numerical Solution of Estimating Equations

Our computational strategy involves two loops: the outer loop searches over the variance component parameter (ξ_1,ξ_2,ξ_3), and the inner loop solves the quasi-score equations for the fixed effects given the current value of (ξ_1,ξ_2,ξ_3).

Given the variance component estimate, $(\hat{\xi}_1,\hat{\xi}_2,\hat{\xi}_3)$, the resulting $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}_t^2$ can be obtained by solving the first two equations in system (34). For the Gaussian ATOMM, the $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}_t^2$ have closed-form expression. For the Binomial-like ATOMM, the score equation for $\boldsymbol{\beta}$ is itself a set of nonlinear equations where nonlinearity originates from the fact that $\boldsymbol{M}$ and $\boldsymbol{\mu}$ are functions of $\boldsymbol{\beta}$. In this case, we use a modified Newton-Raphson algorithm with Fisher scoring [15] to find a convergence value of $\hat{\boldsymbol{\beta}}$. At each step of the Newton-Raphson algorithm, $\hat{\boldsymbol{\beta}}$ is iteratively updated by

$$\hat{\boldsymbol{\beta}}^{(t+1)}=\hat{\boldsymbol{\beta}}^{(t)}+\left(\boldsymbol{X}^T\boldsymbol{M}\boldsymbol{\Sigma}^{-1}\boldsymbol{M}\boldsymbol{X}\right)^{-1}\left[\boldsymbol{X}^T\boldsymbol{M}\boldsymbol{\Sigma}^{-1}\boldsymbol{M}^{-1}(\boldsymbol{Y}-k\boldsymbol{\mu})\right]\big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}^{(t)}}.$$

Once a limiting value of $\hat{\boldsymbol{\beta}}$ is obtained, we plug it into the second equation of (34) and solve to obtain a closed-form expression for $\hat{\sigma}_t^2$, and then plug both $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}_t^2$ into the last three equations of (34).

In the outer loop, we search for the value of the variance component parameter, $(\hat{\xi}_1,\hat{\xi}_2,\hat{\xi}_3)$, that solves the last three equations in (34). We define an objective function, $f(\xi_1,\xi_2,\xi_3)$, to be the total sum of the absolute difference between two sides of the last three equations in (34). The function $f(\xi_1,\xi_2,\xi_3)$ is then to be minimized in the simplex $\{(\xi_1,\xi_2,\xi_3):0\le\xi_i\le 1 \text{ for } i=1,2,3, \text{ and } 0\le\xi_1+\xi_2+\xi_3\le 1\}$ using the Nelder-Mead simplex algorithm [18]. Heuristically, we choose to restart the simplex search twice in order to obtain a better convergence behavior. The search is stopped when either (a) the reduction in the objective function is less than 10^{-2}, or (b) the number of updates exceeds 5,000, whichever occurs first. We denote the output by $(\hat{\xi}_{1,0},\hat{\xi}_{2,0},\hat{\xi}_{3,0})$. Note that for the Gaussian ATOMM, $(\hat{\xi}_{1,0},\hat{\xi}_{2,0},\hat{\xi}_{3,0})$ actually represents the profile MLE of (ξ_1,ξ_2,ξ_3) under the null hypothesis.

3 Score Tests for Assessing Genetic Effects

For association studies with two organisms, there are several hypothesis tests of interest depending on the specific goal. Given a pair of genetic variants, one from the host genome (call this variant H) and the other from the pathogen genome (call this variant P), ATOMM has the option to test for (a) marginal effect of H or P, (b) gene × gene interaction between H and P, (c) effect of H or P allowing for interaction between H and P, and (d) joint effect of H and P. We present a series of score tests based on the binomial ATOMM (31) and (32), where the analogy for Gaussian ATOMM (30) follows naturally.

3.1 Marginal Effect of H or P

One can screen the host or pathogen genome to assess the marginal effect of each individual variant. The marginal association test (e.g., for a host variant) is defined as

$$\mathcal{H}_0 \colon \gamma_1 = 0, \quad \text{versus} \quad \mathcal{H}_A \colon \gamma_1 \neq 0,$$

with constraint $\gamma_2 = \gamma_3 = 0$. The corresponding score statistic is

$$\begin{aligned} T_{\text{marginal}} = & \frac{1}{\hat{\sigma}^2_{t,0}} (\boldsymbol{Y} - k\hat{\boldsymbol{\mu}}_0)^T \hat{\boldsymbol{M}}_0^{-1} \hat{\boldsymbol{\Sigma}}_0^{-1} \hat{\boldsymbol{M}}_0 \boldsymbol{G} \\ & \left[\boldsymbol{G}^T \hat{\boldsymbol{M}}_0 \hat{\boldsymbol{\Sigma}}_0^{-1} \hat{\boldsymbol{M}}_0 \boldsymbol{G} - \boldsymbol{G}^T \hat{\boldsymbol{M}}_0 \hat{\boldsymbol{\Sigma}}_0^{-1} \hat{\boldsymbol{M}}_0 \boldsymbol{W} (\boldsymbol{W}^T \hat{\boldsymbol{M}}_0 \hat{\boldsymbol{\Sigma}}_0^{-1} \hat{\boldsymbol{M}}_0 \boldsymbol{W})^{-1} \boldsymbol{W}^T \hat{\boldsymbol{M}}_0 \hat{\boldsymbol{\Sigma}}_0^{-1} \hat{\boldsymbol{M}}_0 \boldsymbol{G} \right]^{-1} \\ & \boldsymbol{G}^T \boldsymbol{M}_0 \hat{\boldsymbol{\Sigma}}_0^{-1} \boldsymbol{M}_0^{-1} (\boldsymbol{Y} - k\hat{\boldsymbol{\mu}}_0), \end{aligned} \tag{35}$$

where in (35) we set $\boldsymbol{G} = \boldsymbol{G}^{h,\text{test}} \otimes \boldsymbol{1}_{n_p}$, $\boldsymbol{W} = \boldsymbol{X}$ with $\boldsymbol{X}$ being covariates, and the quantities with subscript "0" denote the null estimates obtained by plugging in the null estimates, $\hat{\boldsymbol{\beta}}_0$, $\hat{\sigma}^2_{t,0}$, $\hat{\xi}_{1,0}$, $\hat{\xi}_{2,0}$, $\hat{\xi}_{3,0}$ described in Section 2. Under the null, T_{marginal} follows a χ^2_1 distribution. The marginal association of a pathogen variant can be assessed similarly, in which we test the null $\mathcal{H}_0 \colon \gamma_2 = 0$ against $\mathcal{H}_A \colon \gamma_2 \neq 0$ with constraint $\gamma_1 = \gamma_3 = 0$. The form of marginal association test is similar to that derived from a standard single-organism LMM [1, 13] except that in our case we account for the polygenic effects in the mixed model arising for both organisms as well as their interaction.

3.2 Joint Effect of H and P

A joint test serves as a useful tool to detect overall association due to either the host or pathogen main effects or their interaction. Instead of beginning with a scan of main effects as in typical GWAS, one could, in principle, begin with a scan of all possible host-pathogen SNP-pair effects using the joint tests, for which the null hypothesis is

$$\begin{gathered} \mathcal{H}_0 \colon \gamma_1 = \gamma_2 = \gamma_3 = 0, \\ \text{versus} \\ \mathcal{H}_A \colon \gamma_i \neq 0 \text{ for some } i. \end{gathered} \tag{36}$$

The corresponding test statistic is the same as in (35), except that we set $\boldsymbol{G} = (\boldsymbol{G}^{h,\text{test}} \otimes \boldsymbol{1}_{n_p},\ \boldsymbol{1}_{n_h} \otimes \boldsymbol{G}^{p,\text{test}},\ \boldsymbol{G}^{h,\text{test}} \otimes \boldsymbol{G}^{p,\text{test}})$ and $\boldsymbol{W} = \boldsymbol{X}$. By definition, $\boldsymbol{G}$ here is either an $n \times 3$ or an $n \times 5$ matrix, depending on

whether the variant P is biallelic (i.e., $\boldsymbol{G}^{p,\text{test}} \in \{0,1\}^n$) or triallelic (i.e., $\boldsymbol{G}^{p,\text{test}} \in \{0,1,D\}^n$). The joint test (36) is built upon the same null model as the marginal test, so the estimates, $\hat{\boldsymbol{\beta}}_0$, $\hat{\sigma}_{t,0}$, $\hat{\xi}_{1,0}$, $\hat{\xi}_{2,0}$ and $\hat{\xi}_{3,0}$, remain the same as in Section 3.1. Under $\mathcal{H}_0$, T_{joint} follows either χ^2_3 or χ^2_5, respectively.

The joint test could, in some cases, lead to increased flexibility and power to detect association signal when the variant pair exhibits negligible marginal effects, but strong joint effects. On the other hand, power could be severely compromised by the multiple comparison penalty for the large number of hypothesis tests and by the spending of extra degrees of freedom to test interaction effects in cases when the most important effects are marginal effects.

3.3 Gene × Gene Interaction Between H and P

To identify whether the additive main effects of the variant pair are modified by an additional interaction, a G × G interaction test, for which the null hypothesis is

$$\mathcal{H}_0 \colon \gamma_3 = 0, \quad \text{versus} \quad \mathcal{H}_A \colon \gamma_3 \neq 0.$$

The test statistic is the same as in (35), except that we set $\boldsymbol{G} = \boldsymbol{G}^{h,\text{test}} \otimes \boldsymbol{G}^{p,\text{test}}$, $\boldsymbol{W} = (\boldsymbol{X},\ \boldsymbol{G}^{h,\text{test}} \otimes \mathbf{1}_{n_p},\ \mathbf{1}_{n_h} \otimes \boldsymbol{G}^{p,\text{test}})$, and the estimates, $\hat{\boldsymbol{\beta}}_0$, $\hat{\sigma}^2_{t,0}$, $\hat{\boldsymbol{\Sigma}}_0^{-1}$, $\hat{\boldsymbol{\mu}}_0$, $\hat{\boldsymbol{M}}_0$, are recalculated under $\mathcal{H}_0 \colon \gamma_3 = 0$, instead of $\gamma_1 = \gamma_2 = \gamma_3 = 0$, as described in Section 2. The resulting test statistic has a χ^2_1 or χ^2_2 null distribution depending on whether the variant P is biallelic or triallelic. In principle, for this particular score test, one needs to refit the null variance component, (ξ_1, ξ_2, ξ_3), for each host-pathogen variant pair. However, considering that most variant pairs have only small genetic effects, we choose to instead compute $(\hat{\xi}_1, \hat{\xi}_2, \hat{\xi}_3)$ under the global null (36) only once per genome screen (at least at the initial stage of analysis). The fixed effects $\boldsymbol{\beta}$ and total variance σ^2_t are refit for every host-pathogen variant pair though.

The ability to test the interaction effect separately from the main effects can be particularly useful in the *A. thaliana–X. arboricola* data analysis. The interaction test reveals how a particular host variant responds differently for different pathogen variants. Furthermore, compared to the joint test, the interaction test has fewer degrees of freedom, thereby retaining more statistical power.

3.4 Effect of H or P Allowing for Interaction Between H and P

To assess the genetic association of a given (say, host) variant, one can alternatively jointly test the marginal effect and its interaction with a pathogen variant. This test might help to identify genes that would not be identified by a standard marginal test. More precisely, we focus on

$$\mathcal{H}_0 \colon \gamma_1 = \gamma_3 = 0, \quad \text{versus} \quad \mathcal{H}_A \colon \gamma_1 \neq 0 \text{ or } \gamma_3 \neq 0, \tag{37}$$

without constraint on γ_2. The derived statistic is similar to (35) where we now set $\boldsymbol{W} = (\boldsymbol{X},\ \mathbf{1}_{n_h} \otimes \boldsymbol{G}^{p,\text{test}})$, $\boldsymbol{G} = \big(\boldsymbol{G}^{h,\text{test}} \otimes \mathbf{1}_{n_p},\ \boldsymbol{G}^{h,\text{test}} \otimes \boldsymbol{G}^{p,\text{test}}\big)$, and recalculate the estimates, $\hat{\boldsymbol{\beta}}_0$, $\hat{\sigma}^2_{t,0}$, $\hat{\boldsymbol{\Sigma}}_0^{-1}$, $\hat{\boldsymbol{\mu}}_0$, $\hat{\boldsymbol{M}}_0$, under the null $\mathcal{H}_0 \colon \gamma_1 = \gamma_3 = 0$. To reduce the computational burden, we choose to compute the variance component, $(\hat{\xi}_1, \hat{\xi}_2, \hat{\xi}_3)$, under the global null (36) once per genome scan.

Equation (37) treats the pathogen marginal effect as a nuisance parameter, and treats both the host marginal

effect and the interaction effect as the parameters of interest. If a host variant influences the phenotype through interaction with a given pathogen variant, then allowing for the interaction in the association effect could increase the power to detect the association signal at this host variant.

3.5 Retrospective Score Tests For the Binomial-like Trait

In the *A. thaliana–X. arboricola* data analysis, we found that the marginal p-values from the Gaussian model are well-calibrated, whereas the marginal p-values from the prospective binomial-like model exhibit modest genome-wide inflation (*SI Appendix, Fig. S10*). Note that the test statistic (35) from the binomial-like mixed model is constructed by assessing the null variance of the score function, $(\boldsymbol{Y}-k\hat{\boldsymbol{\mu}}_0)^T\hat{\boldsymbol{M}}_0^{-1}\hat{\boldsymbol{\Sigma}}_0^{-1}\hat{\boldsymbol{M}}_0\boldsymbol{G}$, prospectively, i.e., treating the phenotype $\boldsymbol{Y}$ as random. Furthermore, the variance of $\boldsymbol{Y}$ relies on the mean via the binomial-like variance function. Hence the resulting test might be moderately sensitive to misspecification of the mean model of $\boldsymbol{Y}$ such as omitting some important explanatory variables, and this could be a possible explanation for the modest genome-wide inflation. We refer to the test statistic constructed by modeling the phenotype $\boldsymbol{Y}$ as random as the "prospective" test.

To overcome the apparent lack of robustness of the prospective test under the Binomial-like model, we construct a corresponding retrospective score test, in which we model the (say, host) genotype $\boldsymbol{G}^{h,\text{test}}$ as random, conditional on the phenotype $\boldsymbol{Y}$ and the covariates $\boldsymbol{X}$. Specifically, following our earlier work [12, 22], we use the retrospective statistic,

$$T^{\text{retro}}_{\text{marginal}} = \frac{1}{\sigma_g^2}(\boldsymbol{Y}-k\hat{\boldsymbol{\mu}}_0)^T\hat{\boldsymbol{M}}_0^{-1}\hat{\boldsymbol{\Sigma}}_0^{-1}\hat{\boldsymbol{M}}_0\boldsymbol{G}^{h,\text{test}}\left[(\boldsymbol{Y}-k\hat{\boldsymbol{\mu}}_0)^T\hat{\boldsymbol{M}}_0^{-1}\hat{\boldsymbol{\Sigma}}_0^{-1}\hat{\boldsymbol{M}}_0\hat{\boldsymbol{K}}_h\hat{\boldsymbol{M}}_0\hat{\boldsymbol{\Sigma}}_0^{-1}\hat{\boldsymbol{M}}_0^{-1}(\boldsymbol{Y}-k\hat{\boldsymbol{\mu}}_0)\right]^{-1} \left(\boldsymbol{G}^{h,\text{test}}\right)^T\boldsymbol{M}_0\hat{\boldsymbol{\Sigma}}_0^{-1}\boldsymbol{M}_0^{-1}(\boldsymbol{Y}-k\hat{\boldsymbol{\mu}}_0), \tag{38}$$

to test for association under the Binomial-like model for a host variant with genotype $\boldsymbol{G}^{h,\text{test}}$. Under the null hypothesis (35), $T^{\text{retro}}_{\text{marginal}}$ follows a χ^2_1 distribution. The statistic (38) can be viewed as assessing the null variance of the score function, $(\boldsymbol{Y}-k\hat{\boldsymbol{\mu}}_0)^T\hat{\boldsymbol{M}}_0^{-1}\hat{\boldsymbol{\Sigma}}_0^{-1}\hat{\boldsymbol{M}}_0\boldsymbol{G}^{h,\text{test}}$, in a retrospective manner based on the following population genetic model for biallelic $\boldsymbol{G}^{h,\text{test}}$ under the null:

$$\begin{aligned}\text{E}_0(\boldsymbol{G}^{h,\text{test}}|\boldsymbol{Y},\boldsymbol{X}) &= \alpha\boldsymbol{X},\\ \text{Var}_0(\boldsymbol{G}^{h,\text{test}}|\boldsymbol{Y},\boldsymbol{X}) &= \sigma_g^2\boldsymbol{K}_h.\end{aligned}$$

We could apply an analogous retrospective test to the pathogen instead of the host.

In the case of an inbred line or haploid organism, we consider two possible estimators of σ_g^2. The first is given by a generalization of the sample variance [20]:

$$\hat{\sigma}_g^2 = (n_h-1)^{-1}\left((\boldsymbol{G}^{h,\text{test}})^T\hat{\boldsymbol{K}}_h^{-1}\boldsymbol{G}^{h,\text{test}} - (\boldsymbol{G}^{h,\text{test}})^T\hat{\boldsymbol{K}}_h^{-1}\boldsymbol{1}_{n_h}(\boldsymbol{1}_{n_h}^T\hat{\boldsymbol{K}}_a^{-1}\boldsymbol{1}_{n_h})^{-1}\boldsymbol{1}_{n_h}^T\hat{\boldsymbol{K}}_h^{-1}\boldsymbol{G}^{h,\text{test}}\right), \tag{39}$$

if the matrix $\hat{\boldsymbol{K}}_h$ is non-singular. If $\hat{\boldsymbol{K}}_h$ is singular, for example, in the case of no missing genotypes at any variants being considered in the estimator (5) (with $\hat{f}_l$ used in place of f_l), we take $\hat{\sigma}_g^2 = (n_h-1)^{-1}(\boldsymbol{G}^{h,\text{test}})^T\hat{\boldsymbol{K}}_h^{-}\boldsymbol{G}^{h,\text{test}}$, where $\hat{\boldsymbol{K}}_h^{-}$ is the Moore-Penrose generalize inverse. See [20] for details. The second estimator of σ_g^2 we consider is given by

$$\check{\sigma}_g^2 = \check{p}(1-\check{p}), \quad \text{where} \quad \check{p} = (\boldsymbol{1}_{n_h}^T\hat{\boldsymbol{K}}_h^{-1}\boldsymbol{1}_{n_h})^{-1}\boldsymbol{1}_{n_h}^T\hat{\boldsymbol{K}}_h^{-1}\boldsymbol{G}^{h,\text{test}}. \tag{40}$$

(Note that in both estimators of σ_g^2, we implicitly assume $\mathbb{E}_0(\boldsymbol{G}^{h,\text{test}}|\boldsymbol{Y},\boldsymbol{X}) = p\mathbf{1}$ instead of $\alpha\boldsymbol{X}$.) In diploid organisms, the analogue of estimator $\hat{\sigma}_g^2$ in equation (39) is recommended in preference to analogue of estimator $\check{\sigma}_g^2$ because the analogue of $\hat{\sigma}_g^2$ does not require Hardy-Weinberg equilibrium. However, in an inbred line or haploid organism, no such consideration applies, and the general relationship between mean and variance of a binary random variable, given by $\sigma^2 = \mu(1-\mu)$, necessarily holds. The estimator $\hat{\sigma}_g^2$ does not make use of this information. In our data analysis, when we perform retrospective association tests on variants in the *X. arboricola* genome, we find that $\hat{\sigma}_g^2$ is numerically unstable for several *X. arboricola* variants (*SI Appendix, Fig. S11*). This is probably due to the relatively small sample size ($n_p = 22$), and the fact that we have some strain pairs with $\hat{K}_h(i,j)$ close to 1, so for SNPs for which one of these strain pairs does not agree, $\hat{\sigma}_g^2$ becomes highly inflated. In contrast, $\check{\sigma}_g^2$ continues to perform reasonably in this context. When testing a triallelic variant, we use a version of equations (38) and (40) extended to multi-allelic variants (see [11] for details).

4 Characterization of *Xanthomonas* Strains

4.1 Isolation of *Xanthomonas* Strains from Natural Populations of *A. thaliana*

The 12 US strains of *Xanthomonas* were isolated from three locations (Lake Michigan College, LMC, latitude = 42°5′24.41″N, longitude = 86°23′36.27″W; Michigan Extension, ME & MEDV, latitude = 42°5′33.72″N, longitude = 86°21′22.76″W; North Liberty, NL, latitude = 41°32′24.88″N, longitude = 86°25′32.86″W). The 12 French strains of *Xanthomonas* were isolated from four locations (Brendaouez, BRE, latitude = 48°36.909′N, longitude = 4°25.129′WLa Forest Landerneau, FOR, latitude = 48°25.560′N, longitude = 4°18.413′W; Meurchin, MEU, latitude = 50°29.947′N, longitude = 2°53.633′E; Ploudiry, PLY, latitude = 48°27.502′N, longitude = 4°8.383′W). The 24 strains of *Xanthomonas* were first identified based on the colony morphology. The 12 US strains were then sequenced for a 16S fragment with the 799f primer, whereas the 12 French strains were sequenced for a *rpoD* fragment [7].

4.2 Phylogenetic Analysis

To identify the phylogenetic position of the 24 *Xanthomonas* strains in our study, we first built a phylogeny based on the *rpoD* housekeeping gene by using 76 *Xanthomonas* strains whose genomes were available in GenBank. A phylogenetic tree showed that the 24 studied strains belong to the *X. arboricola* complex (*SI Appendix, Fig. S6*). To verify whether the phylogenic relationship of the *Xanthomonas* strains was robust, we also performed a MLST phylogeny based on *rpoD*, *gyrB*, *atpD*, and *glnA* housekeeping genes (5865 bp). MLST also supported that the 24 strains belong to the *X. arboricola* complex (*SI Appendix, Fig. S7*).

4.3 Testing the Pathogenicity of *X. arboricola*

Because *X. arboricola* has been mainly described as a bacterial pathogen of woody plants, we tested whether the 24 studied strains were pathogenic on *A. thaliana*. To do so, we inoculated four lines of *A. thaliana* (Col-0, Kas-1, Pro-0 and VED-10) with: (i) four strains isolated from walnut tree [4], with the strains CFBP2528 and CFBP7179 described as pathogenic on walnut and the strains CFBP7634 and CFBP7651

reported to be non-pathogenic and to lack the T3SS; (ii) four strains used in our study (two strains lacking the *hrp/hrc* cluster, LMC_P73 and FOR_F23, and two strains with the *hrp/hrc* cluster, MEDV_37 and MEU_M1). Following the protocol described in [10], our pathogenic assay revealed that all the strains tested were pathogenic on *A. thaliana*, and that significant phenotypic variation in the response to *X. arboricola* was found among the four *A. thaliana* lines (*SI Appendix, Fig. S8*).

4.4 DNA Extraction, Genome Sequencing and Bioinformatics Analysis

DNA of the *X. arboricola* strains was extracted according to the Qiagen Gentra Puregene protocol for gram-negative bacteria. The only modification was to keep the samples on ice for the protein precipitation for 2 hours instead of 5 minutes. DNAseq was performed at the Argonne National Laboratory. DNAseq libraries were prepared according to Illumina's protocol using the Illumina's TruSeq DNAseq Library Preparation and Illumina Mate Pair Library Preparation kits. DNAseq experiments were performed on an Illumina HiSeq2000 using a paired-end read length of 2 $\times$ 150 pb. The *X. arboricola* sequences were assembled using A5 pipeline [21] and aligned using progressiveMauve [6]. SNPs were compiled from the progressiveMauve alignments using a custom Python script.

Structural annotation for all *X. arboricola* assembled genomes was performed by using the EuGene-PP pipeline [19] in a Galaxy environment [5]. Functional annotation was conducted after converting the assembled genomes in proteomes with InterPro algorithm [16]. OrthoMCL analysis [14] was performed to cluster the ortholog and paralog protein families by using 80% as match cut-off in ortholog clustering. Statistics for the annotated genomes are shown in the *SI Appendix, Table S4*.

4.5 Comparative Genome Analysis of *X. arboricola* and Analysis of the Effector Repertoire Composition

Eight strains isolated from crops (CFBP2528, CFBP7179, CFBP7634, CFBP7651, CITA44, IVIA2626.1, NCPPB1832 and NCPPB1630) were also selected for comparative genomics analysis because of their close relationship with the 24 studied strains.

To characterize the type III secretion system (T3SS) repertoire, a list of protein sequences for the *hrp/hrc* cluster and the T3SS effectors were established by using the bacterial type III secretion system database previously published `https://biocomputer.bio.cuhk.edu.hk/T3DB/` [23] and the effectors sequences described in [9] and [24]. The protein sequence databases for both effectors and *hrp/hrc* cluster (*Dataset1* and *Dataset2*) were blasted against each genome by using the NCBI BLAST + suite.makeblastdb command line tool that creates a BLAST database for several FASTA files. Detailed information about this process is available at `http://www.ncbi.nlm.nih.gov/books/NBK279690/`. We assigned an effector to a given allele present in the database when: i) its protein sequence could be aligned for over 80% of the length of the sequence present in the database, and ii) it displayed a homology level $>$ 85%. Sequences for the same effector were considered as different allelic forms if their homology identity was $<$ 90%.

The 24 strains isolated from *A. thaliana* and the eight crop strains listed above were analyzed for their effector repertoire composition and for the presence of the *hrp/hrc* cluster. Firstly, only three strains (MEDV_A37, MEU_M1 and NL P_126) isolated from *A. thaliana* carried a *hrp/hrc* cluster (*SI Appendix, Table S2*). Among

the crop strains, results were coherent with those already published in the literature. Strains CFBP 7634 and CITA 44 lacked the *hrp/hrc* cluster [4, 8] and strains CFBP 2528, CFBP 7651, CFBP 7179, NCPPB1630 and IVIA2626.1 carried the *hrp/hrc* cluster (*SI Appendix, Table S2*). Secondly, T3SS effector distribution was related with the presence/absence of the *hrp/hrc* cluster. Only the strains lacking the *hrp/hrc* cluster carry the avirulence genes *AvrXccA1* and *AvrXccA2* (*SI Appendix, Table S3*). On the other hand, *hrp/hrc*-positive strains contain the *AvrBs2* effector and displayed a variable effector composition. Natural strains NL_P126 and MEDV_A37 had six and four T3SS effectors respectively, while the strain MEU_M1 only carried two effectors (*SI Appendix, Table S3*). Finally, strains isolated from crop plants were those with the higher number of effector genes (*SI Appendix, Table S3*).

5 Phenotypic Models for QDR in *A. thaliana–X.arboricola* Study

In this section, we provide the results for a variety of other phenotypic models for QDR. Such models serve as exploratory data analyses before we perform association analyses. For ease of comparison, we adjust for person effects by regressing them out prior to the analysis. In addition to the ATOMM model applied to the full data set, we consider the following five models: (1) a fixed-effects model that includes indicators for each of the 22 *X. arboricola* strains, the 130 *A. thaliana* lines, and their interactions as predictors; (2) a standard random-effects model with GRMs $\boldsymbol{K}_h$ and $\boldsymbol{K}_p$ in (30) replaced by identity matrices; (3) ATOMM applied to the subset of the data obtained by removing seven strains with minimal marginal effects (*SI Appendix, Figure S3*, 6 US strains and 1 France strain FOR_F26); (4) a fixed-effects model obtained by grouping similar *X. arboricola* strains together; and (5) ATOMM with a pathogen country of origin indicator as a fixed effect. We included plant random effects in all of the above models. The details of the model formulations considered are described below.

5.1 Model Formulation

1. Fixed-effects model:

$$\begin{aligned} &\boldsymbol{Y}|\text{host-pathogen pairs} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad \text{where} \\ &\boldsymbol{\mu} = \beta_0 \mathbf{1} + \sum_{i=1}^{129} \lambda_i \mathbb{1}_{\text{host } i} + \sum_{j=1}^{21} \beta_j \mathbb{1}_{\text{pathogen } j} + \sum_{i=1}^{129}\sum_{j=1}^{21} \eta_{ij} \mathbb{1}_{\text{host } i \text{ and pathogen } j}, \\ &\boldsymbol{\Sigma} = \sigma_j^2 \mathbf{J} + \sigma_e^2 \mathbf{I}, \end{aligned} \tag{41}$$

   where λ_i is the marginal effect of host line i, β_j is the marginal effect of pathogen strain j, η_{ij} denotes the interaction effect between host line i and pathogen strain j, $\mathbf{J}$ is a covariance matrix with $\boldsymbol{J}_{k\ell} = 1$ if k and ℓ represents two leaves from the same plant, and 0 otherwise, σ_j^2 is variance of the plant random effect, and σ_e^2 is the variance of i.i.d. environmental noise.

2. i.i.d. Random-effects model:

$$\begin{aligned} &\boldsymbol{Y}|\text{host-pathogen pairs} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma_t^2 \boldsymbol{\Sigma}), \quad \text{where} \\ &\boldsymbol{\mu} = \beta_0 \mathbf{1}, \end{aligned} \tag{42}$$

$$\boldsymbol{\Sigma} = \xi_h \boldsymbol{Z}_h \boldsymbol{Z}_h^T + \xi_p \boldsymbol{Z}_p \boldsymbol{Z}_p^T + \xi_{hp} \boldsymbol{Z}_{hp} \boldsymbol{Z}_{hp}^T + \xi_J \boldsymbol{J} + (1 - \xi_h - \xi_p - \xi_{hp} - \xi_J)\mathbf{I},$$

where $\boldsymbol{Z}_h$, $\boldsymbol{Z}_p$, and $\boldsymbol{Z}_{hp}$ are the incidence matrices that map the observed QDR to host lines, to pathogen strains, and to host-pathogen pairs, respectively; $\boldsymbol{J}$ is the same covariance matrix as before, σ_t^2 is the total residual variance, and ξ_h, ξ_p, ξ_{hp}, ξ_J represent the proportion of the variance explained by host, pathogen, host-pathogen and plant i.i.d. random effects, respectively.

3. ATOMM null model:

$$\begin{aligned} &\boldsymbol{Y}|\text{host-pathogen pairs} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma_t^2 \boldsymbol{\Sigma}), \quad \text{where} \\ &\boldsymbol{\mu} = \beta_0 \mathbf{1}, \\ &\boldsymbol{\Sigma} = \xi_h \boldsymbol{Z}_h \boldsymbol{K}_h \boldsymbol{Z}_h^T + \xi_p \boldsymbol{Z}_p \boldsymbol{K}_p \boldsymbol{Z}_p^T + \xi_{hp} \boldsymbol{Z}_{hp} (\boldsymbol{K}_h \otimes \boldsymbol{K}_p) \boldsymbol{Z}_{hp}^T + \xi_J \boldsymbol{J} + (1 - \xi_h - \xi_p - \xi_{hp} - \xi_J)\mathbf{I}, \end{aligned} \tag{43}$$

where all notation is the same as before, and $\boldsymbol{K}_h$, $\boldsymbol{K}_p$ are GRMs for host and for pathogen (see Sections 1.1.1 and 1.1.2), respectively. In the full analysis, we fit the above model to $n_{\text{subset}} = 32{,}960$ observations. As suggested by a reviewer, we also fit the above model to the subset of the data ($n_{\text{subset}} = 22{,}478$) obtained by considering only the following 15 strains:

- US strains: LMC_P47, LMC_P73, MEDV_P26, LMC_P25, MEDV_A37, MEDV_P39;
- France strains: MEU_M1 , FOR_F21, FOR_F23, PLY_3, PLY_2, FOR_F20, PLY_9, PLY_1, PLY_4.

4. Fixed-effects group membership model:

$$\begin{aligned} &\boldsymbol{Y}|\text{strain membership} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \text{where} \\ &\boldsymbol{\mu} = \sum_{i=1}^{4} \sum_{j=1}^{2} \gamma_{ij} \mathbb{1}_{\text{host group } i \text{ and pathogen group } j}, \\ &\boldsymbol{\Sigma} = \sigma_j^2 \boldsymbol{J} + \sigma_e^2 \boldsymbol{I}, \end{aligned} \tag{44}$$

where all notation is the same as before, γ_{ij} represents the effect on QDR of paring a host from *A. thaliana* group i with a pathogen from *X. arboricola* group j.

There are different ways to group *X. arboricola* strains. One natural choice is to group *X. arboricola* strains into US vs. France groups. Another possibility, as suggested by a reviewer, is to group *X. arboricola* strains based on association peaks in the separate *A. thaliana* GWA mapping (*SI Appendix, Figures S12* and *S13*). However, the separate analyses have poor power to detect association due to their small sample sizes, and we do not observe informative similarity among association peaks that would enable us to assemble strains into groups. Therefore, we choose the first grouping scheme (i.e. US vs. France) and include the group membership as the fixed effect in the model (44) (*Fig 2* and section **Population Structure and Effects** in the main paper).

5. ATOMM + fixed-effects pathogen group membership model:

$$\begin{aligned} &\boldsymbol{Y}|\text{host-pathogen pairs} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma_t^2 \boldsymbol{\Sigma}), \quad \text{where} \\ &\boldsymbol{\mu} = \beta_0 \mathbf{1} + \beta_1 \mathbb{1}_{\text{US strain}}, \\ &\boldsymbol{\Sigma} = \xi_h \boldsymbol{Z}_h \boldsymbol{K}_h \boldsymbol{Z}_h^T + \xi_p \boldsymbol{Z}_p \boldsymbol{K}_p' \boldsymbol{Z}_p^T + \xi_{hp} \boldsymbol{Z}_{hp} (\boldsymbol{K}_h \otimes \boldsymbol{K}_p') \boldsymbol{Z}_{hp}^2 + \xi_J \boldsymbol{J} + (1 - \xi_h - \xi_p - \xi_{hp} - \xi_J)\mathbf{I}, \end{aligned} \tag{45}$$

where all notation is the same as before, and $\boldsymbol{K}_p'$ is the modified pathogen GRM after accounting for pathogen group membership indicators (i.e., we use group-specific MAF in the calculation of the GRM).

5.2 Assessing Misfit of the Model with i.i.d. Random Effects (Model 2)

We consider the problem of assessing the misfit of Models 2 and 3 defined in the previous subsection. Consider a larger model (call it Model 4), that has both Models 2 and 3 as sub-models. Specifically, Model 4 has everything that is in Model 3 (ATOMM), plus additional i.i.d. random line effects, i.i.d. random strain effects, and i.i.d. random interactions. Because Model 2 is a sub-model of Model 4, then in principle, we could assess the goodness of fit of Model 2 by comparing its maximized log-likelihood to that of Model 4. Similarly, because Model 3 is a sub-model of Model 4, we could in principle assess the goodness of fit of Model 3 by comparing its maximized log-likelihood to that Model 4. Let l_i denote the maximized log-likelihood of Model i, $i = 2, 3, 4$. Then since Models 2 and 3 are sub-models of Model 4, it is always true that $l_2 \leq l_4$ and $l_3 \leq l_4$. Furthermore, the goodness of fit test statstic for Model 2 is $T_2 = 2 * (l_4 - l_2) \geq 0$ which follows a χ^2 with 3 degrees of freedom under the null hypothesis that Model 2 is the true model. (Similarly, the goodness of fit test statistic for Model 3 is $T_3 = 2 * (l_4 - l_3) \geq 0$, which follows a χ^2 with 3 degrees of freedom under the null hypothesis that Model 3 is the true model.) A major difficulty is that for our data, Model 4 has 8 variance components, which makes it computationally challenging to fit. However, it turns out that in our data analysis, we could still reject Model 2 with p-value $< 4 \times 10^{-4}$ based only on l_2 and l_3. The reason is that, since T_2 and T_3 are both always nonnegative, $T_2 \geq T_2 - T_3 = 2 * (l_3 - l_2)$, so we can compare $2 * (l_3 - l_2)$ to a χ^2 distribution with 3 degrees of freedom to obtain an upper bound of 4×10^{-4} on the p-value for T_2. (We could, of course, do the mirror image analysis for T_3, except that in our data analysis, $T_3 - T_2 < 0$, so the upper bound on the p-value for T_3 is 1.)

5.3 Model Comparison

In what follows, we refer to the fixed-effects model of equation (41) as Model 1, the i.i.d. random-effects model of equation (42) as Model 2, and the ATOMM model as Model 3, where all three models are fit to the full data.

Of the three models, Model 3 (ATOMM model) is the most preferred by the BIC model selection criterion (*SI Appendix, Table S5*), indicating that it is the most appropriate one to use for statistical inference such as association mapping.

Model 1 (fixed effects model) is of interest because it captures the full genetic effects of host line and pathogen strain. In other words, in Model 1, the proportion of variance explained by host line effects accounts for both additive polygenic and epistatic effects of line; the proportion of variance explained by pathogen strain effects similarly accounts for both additive polygenic and epistatic effects of strain; and the proportion of variance explained by host-pathogen interaction effects similarly accounts for all line-strain genetic interaction effects. (Recall that there are no dominance effects in this pathosystem because the host is a diploid inbred organism and the pathogen is a haploid organism.) Thus, it is potentially interesting to compare the estimated proportions of variance from Model 1 to those from Model 3, in which only additive genetic effects are modeled. One major drawback to Model 1, though, is that with 2,860 fixed effects in

the model (including intercept), there is a severe problem of bias-variance trade-off, with the result that the standard errors for the parameters of interest are quite large, thus limiting the comparison somewhat.

However, in *SI Appendix, Table S5*, it is notable that the proportion of variance explained by the full genetic effects of strain is estimated at 52% in Model 1, while the proportion of variance explained by the additive polygenic effects of strain is estimated at 44% in Model 3, showing that QDR is highly heritable with respect to the pathogen and that nearly all of the strain effect is attributable to additive polygenic effects of variants in the *X. arboricola* genome, so that the responses to related strains tend to be similar.

From the extremely poor value of the BIC model selection criterion for Model 1 relative to the other models (*SI Appendix,Table S5*; lower BIC is better), we can see that Model 1 is vastly over-parameterized, so, e.g., not suitable for use in association mapping. Indeed, when SNP fixed effects are introduced for mapping, only SNPs whose genotypes are orthogonal to the set of fixed covariates could possibly show any association, so in Model 1, mapping would be impossible because any genetic variation to be tested would be completely confounded with the fixed effects already included.

Model 2 seems to be of less interest because it forces all the line and strain effects to be non-genetic, in the sense that, e.g., all strain effects are forced to be i.i.d. with a common variance, with all of the strain effects uncorrelated, regardless of the fact that some pairs of strains are $> 90\%$ identical. In fact, Model 2 shows significant misfit to the data (p-value $< .0004$; see section 5.2 for details). Thus, the parameter estimates under this model may not be very meaningful.

We also add a sub-analysis (suggested by a reviewer) for a subset of 15 strains (*SI Appendix, Table S6*), consisting of 7 US strains and 8 France strains, detailed in subsection 5.1. We fit the same three models described above: Model 1 with fixed effects for each line, each strain, and each line-strain pair, Model 2 with i.i.d random effects where all GRMs are set to the identity matrix, and Model 3 which is the ATOMM model. The results are similar to those from the full data set, except that compared to the full data set, the sub-analysis exhibits smaller inter-strain variation, which is obviously expected because of the way the strains were chosen — the ones with lowest values were removed, thus, automatically reducing the variation.

In the fixed-effects model, there are different ways to group *X. arboricola* strains. One natural choice is to divide them into US vs. France groups, which is the grouping we have used in Figure 2. Another possibility, as suggested by a reviewer, is to group *X. arboricola* strains based on association peaks in the separate A. thaliana GWA mapping (*SI Appendix, Figures S12 and S13*). However, the separate analyses have poor power to detect associations due to their small sample sizes, and we do not observe informative similarity among association peaks that would enable us to assemble strains into groups. Therefore, we retain the first grouping scheme (i.e., US vs. France) and include group membership as a fixed effect in the model (Fig 2).

To further investigate the polygenic nature of QDR, we fit the model (equation 45) obtained by including the pathogen group membership indicator as the fixed effect and by modifying the pathogen GRM using group-specific MAFs. We found that the "US strain" effect is negative, consistent with the observation that French strains are more virulent than US strains (*SI Appendix, Table S7*). Furthermore, the *X. arboricola* random polygenic effect still explains a considerable proportion (31.1%) of phenotypic variance after accounting for the population effect, further demonstrating the polygenic nature of QDR (*SI Appendix, Table S7*).

6 Multiple Testing Adjustment for Gene Ontology Analysis

In the **Materials and Methods** section of the paper we described the use of 10,000 permutation replicates to calculate nominal p-values of enrichment for particular BP terms. We were able to use the results of the same 10,000 replicates to also calculate p-values adjusted for multiple comparisons corresponding to the analysis of multiple BP terms, as we now describe.

For a given BP, call it BP b, let O_b denote the Enrichment$_o$ value of BP b, and for BP b and permutation replicate r, let P_{br} denote the Enrichment$_p$ value of BP b in permutation replicate r, where the permutation replicates and the definitions of Enrichment$_o$ and Enrichment$_p$ are as described in **Material and Methods**. Assume that 10,000 permutation replicates have been performed and that P_{br} is available for each BP term in each replicate, while the data include O_b for each BP term. Because the multiple comparisons analysis requires us to be able to make meaningful comparisons across BPs, we analyzed only BPs that had at least 20 hits somewhere among the 10,000 replicates. For BP b, let $m_b = \text{median}(P_{br})$ where the median is taken over all permutation replicates, and let $s_b = 1\%$ trimmed range of P_{br}, where this is also over all permutation replicates r. For each BP in each permutation replicate, we standardize P_{br} by defining $U_{br} = (P_{br} - m_b)/s_b$. Then for each permutation replicate, we calculate $V_r = \max_b U_{br}$. In the data, we calculate $T = \max_b (O_b - m_b)/s_b$. Then we compare T to the empirical distribution of V_r across the 10,000 permutation replicates to obtain the p-value of the relatively most enriched BP (i.e. the BP that maximizes $(O_b - m_b)/s_b$).

References

[1] Mark Abney, Carole Ober, and Mary Sara McPeek. Quantitative-trait homozygosity and association mapping and empirical genomewide significance in large, complex pedigrees: fasting serum-insulin level in the Hutterites. *The American Journal of Human Genetics*, 70(4):920–934, 2002.

[2] Alan Agresti. *Categorical Data Analysis*, volume 990. New York: John Wiley & Sons, 1996.

[3] William Astle and David J Balding. Population structure and cryptic relatedness in genetic association studies. *Statistical Science*, pages 451–471, 2009.

[4] Sophie Cesbron, Martial Briand, Salwa Essakhi, Sophie Gironde, Tristan Boureau, Charles Manceau, Marion Fischer-Le Saux, and Marie-Agnès Jacques. Comparative genomics of pathogenic and non-pathogenic strains of xanthomonas arboricola unveil molecular and evolutionary events linked to pathoadaptation. *Frontiers in Plant Science*, 6, 2015.

[5] Peter JA Cock, Björn A Grüning, Konrad Paszkiewicz, and Leighton Pritchard. Galaxy tools and workflows for sequence analysis with applications in molecular plant pathology. *PeerJ*, 1:e167, 2013.

[6] Aaron E Darling, Bob Mau, and Nicole T Perna. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PloS One*, 5(6):e11147, 2010.

[7] E Fargier, M Fischer-Le Saux, and C Manceau. A multilocus sequence analysis of xanthomonas campestris reveals a complex structure within crucifer-attacking pathovars of this species. *Systematic and applied microbiology*, 34(2):156–165, 2011.

[8] Jerson Garita-Cambronero, Ana Palacio-Bielsa, María M López, and Jaime Cubero. Draft genome sequence for virulent and avirulent strains of xanthomonas arboricola isolated from prunus spp. in spain. *Standards in genomic sciences*, 11(1):12, 2016.

[9] Endrick Guy, Anne Genissel, Ahmed Hajri, Matthieu Chabannes, Perrine David, Sébastien Carrere, Martine Lautier, Brice Roux, Tristan Boureau, Matthieu Arlat, et al. Natural genetic variation of xanthomonas campestris pv. campestris pathogenicity on arabidopsis revealed by association and reverse genetics. *MBio*, 4(3):e00538–12, 2013.

[10] Carine Huard-Chauveau, Laure Perchepied, Marilyne Debieu, Susana Rivas, Thomas Kroj, Ilona Kars, Joy Bergelson, Fabrice Roux, and Dominique Roby. An atypical kinase under balancing selection confers broad-spectrum disease resistance in arabidopsis. *PLoS Genetics*, 9(9):e1003766, 2013.

[11] Johanna Jakobsdottir and Mary Sara McPeek. MASTOR: mixed-model association mapping of quantitative traits in samples with related individuals. *The American Journal of Human Genetics*, 92(5):652–666, 2013.

[12] Duo Jiang, Sheng Zhong, and Mary Sara McPeek. Retrospective binary-trait association test elucidates genetic architecture of crohn disease. *The American Journal of Human Genetics*, 98(2):243–255, 2016.

[13] Hyun Min Kang, Jae Hoon Sul, Susan K Service, Noah A Zaitlen, Sit-yee Kong, Nelson B Freimer, Chiara Sabatti, Eleazar Eskin, et al. Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics*, 42(4):348–354, 2010.

[14] Li Li, Christian J Stoeckert, and David S Roos. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Research*, 13(9):2178–2189, 2003.

[15] Peter McCullagh and James A Nelder. Generalized linear models, No. 37 in monograph on statistics and applied probability, 1989.

[16] Alex Mitchell, Hsin-Yu Chang, Louise Daugherty, Matthew Fraser, Sarah Hunter, Rodrigo Lopez, Craig McAnulla, Conor McMenamin, Gift Nuka, Sebastien Pesseat, et al. The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Research*, pages D213–21, 2014.

[17] Fionn Murtagh. Complexities of hierarchic clustering algorithms: State of the art. *Computational Statistics Quarterly*, 1(2):101–113, 1984.

[18] John A Nelder and Roger Mead. A simplex method for function minimization. *The Computer Journal*, 7(4):308–313, 1965.

[19] Erika Sallet, Jérôme Gouzy, and Thomas Schiex. EuGene-PP: a next generation automated annotation pipeline for prokaryotic genomes. *Bioinformatics*, 30(18):2659–61, 2014.

[20] Timothy Thornton and Mary Sara McPeek. ROADTRIPS: case-control association testing with partially or completely unknown population and pedigree structure. *The American Journal of Human Genetics*, 86(2):172–184, 2010.

[21] Andrew Tritt, Jonathan A Eisen, Marc T Facciotti, and Aaron E Darling. An integrated pipeline for de novo assembly of microbial genomes. *PloS One*, 7(9):e42304, 2012.

[22] Miaoyan Wang, Johanna Jakobsdottir, Albert V Smith, and Mary Sara McPeek. G-strategy: Optimal selection of individuals for sequencing in genetic association studies. *Genetic epidemiology*, 40(6):446–460, 2016.

[23] Yejun Wang, He Huang, Ming'an Sun, Qing Zhang, and Dianjing Guo. T3DB: an integrated database for bacterial type III secretion system. *BMC bioinformatics*, 13(1):66, 2012.

[24] Frank F White, Neha Potnis, Jeffrey B Jones, and Ralf Koebnik. The type III effectors of xanthomonas. *Molecular Plant Pathology*, 10(6):749–766, 2009.

[25] Sheng Zhong, Duo Jiang, and Mary Sara McPeek. CERAMIC: Case-control association testing in samples with related individuals, based on retrospective mixed model analysis with adjustment for covariates. *PLoS Genetics*, 12(10):e1006329, 2016.

Supplementary Figures and Tables

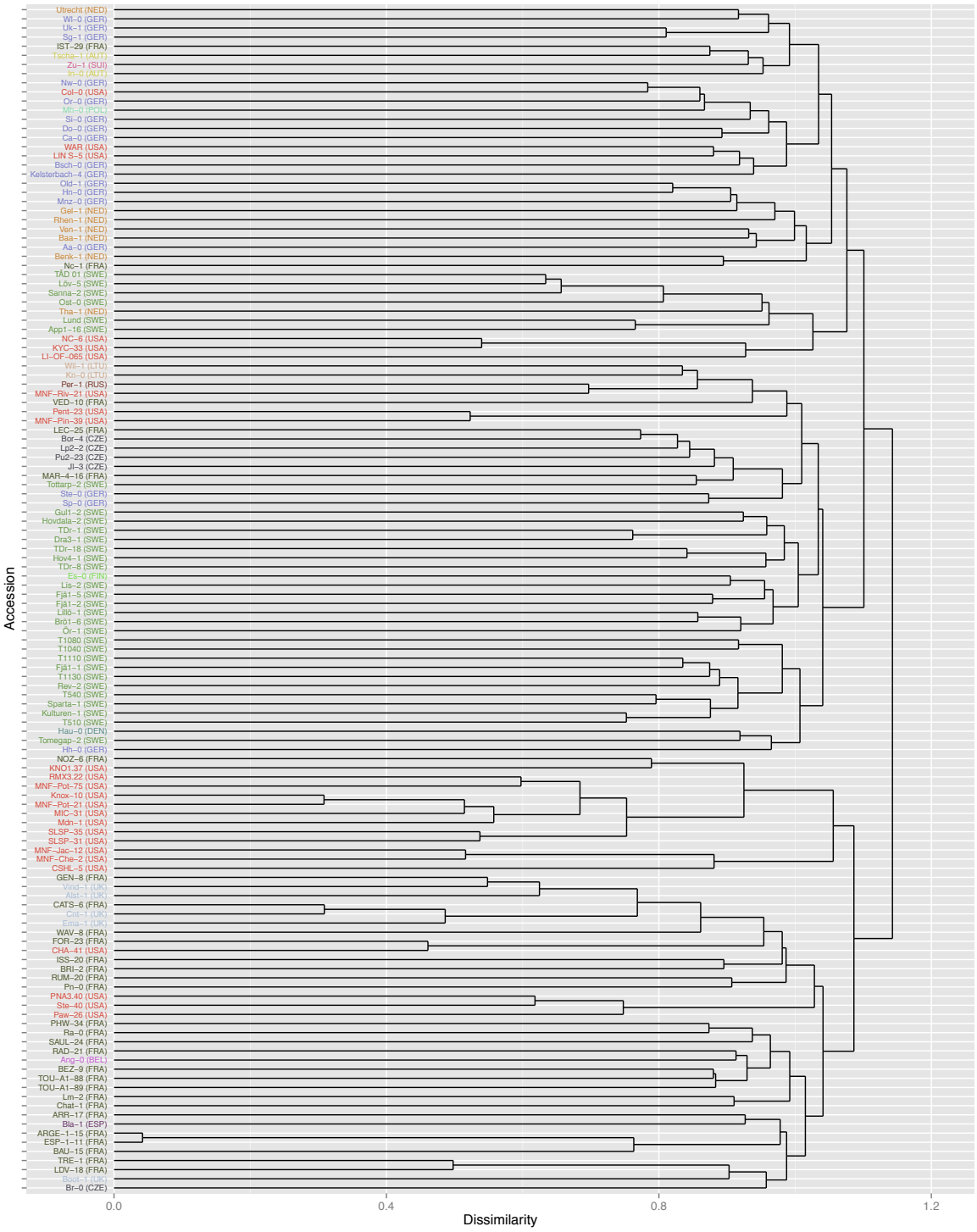

Fig. S1: Hierarchical clustering of the 130 *A. thaliana* lines. We performed hierarchical clustering based on the *A thaliana* GRM using UPGMA [17]. For the dissimilarity measure, we used $d(i,j) = [1 - \hat{\rho}(i,j)]/2$, where $\hat{\rho}(i,j) = \boldsymbol{K}_h(i,j)/\sqrt{\boldsymbol{K}_h(i,i)\boldsymbol{K}_h(j,j)}$ was the estimate of the genomic correlation between *A. thaliana* lines i and j, for all $i, j = 1, \ldots, 130$. We note that the *A. thaliana* lines labeled with the same country of origin tend to be clustered into the same branch.

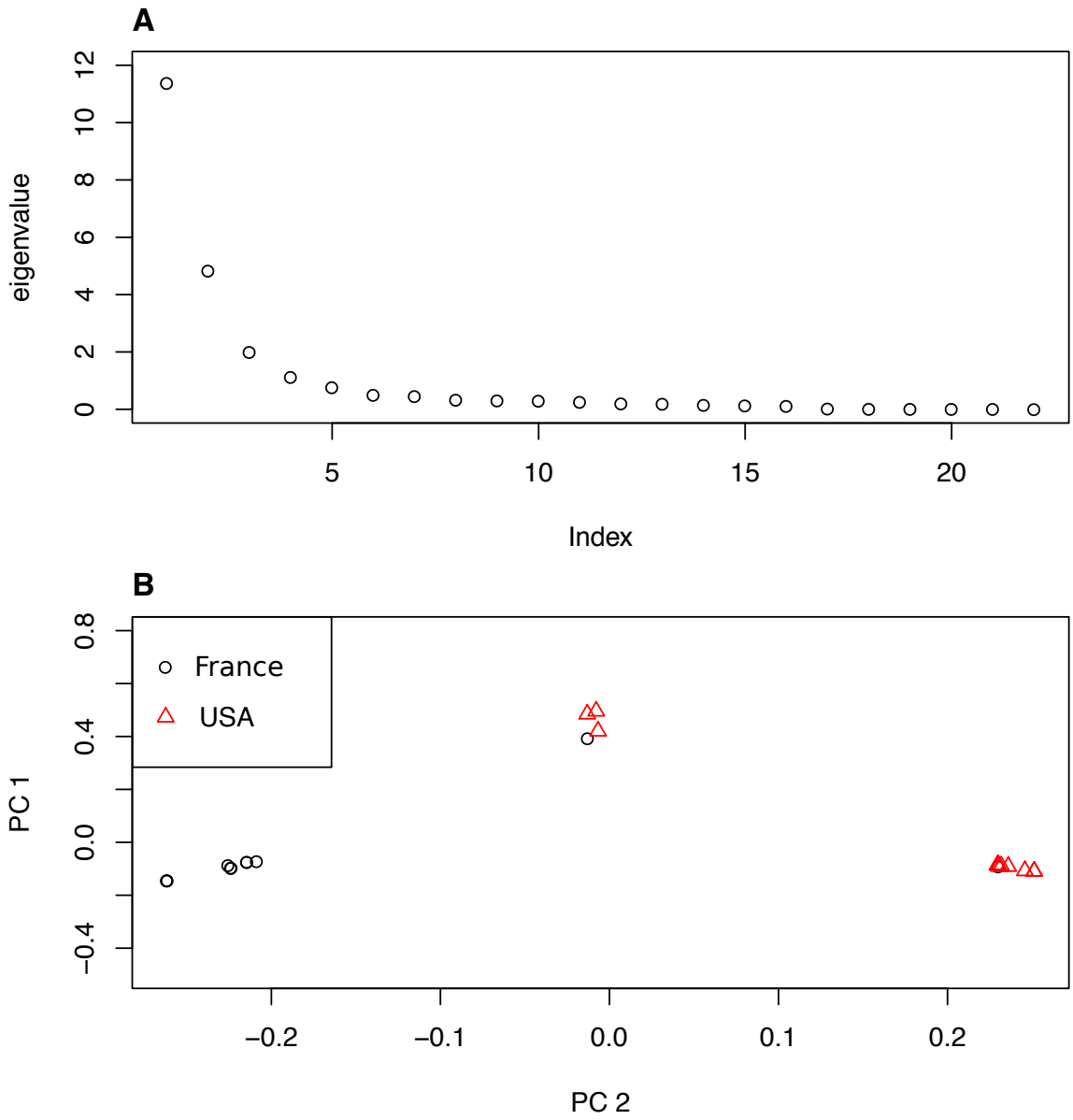

Fig. S2: PCA on the *X. arboricola* GRM. We calculated the *X. arboricola* GRM as proposed in Section 1.1.2 of the *SI Appendix, Supplementary Text*. Panel (A) plots the eigenvalues of the *X. arboricola* GRM in descending order. Panel (B) plots the top two eigenvectors of the *X. arboricola* GRM. The top two eigenvectors clearly separated the US strains (in red) from the France strains (in black), except that the strains BRE_17 and MEU_M1, which originated in France, were more genetically similar to US strains than to the other France strains.

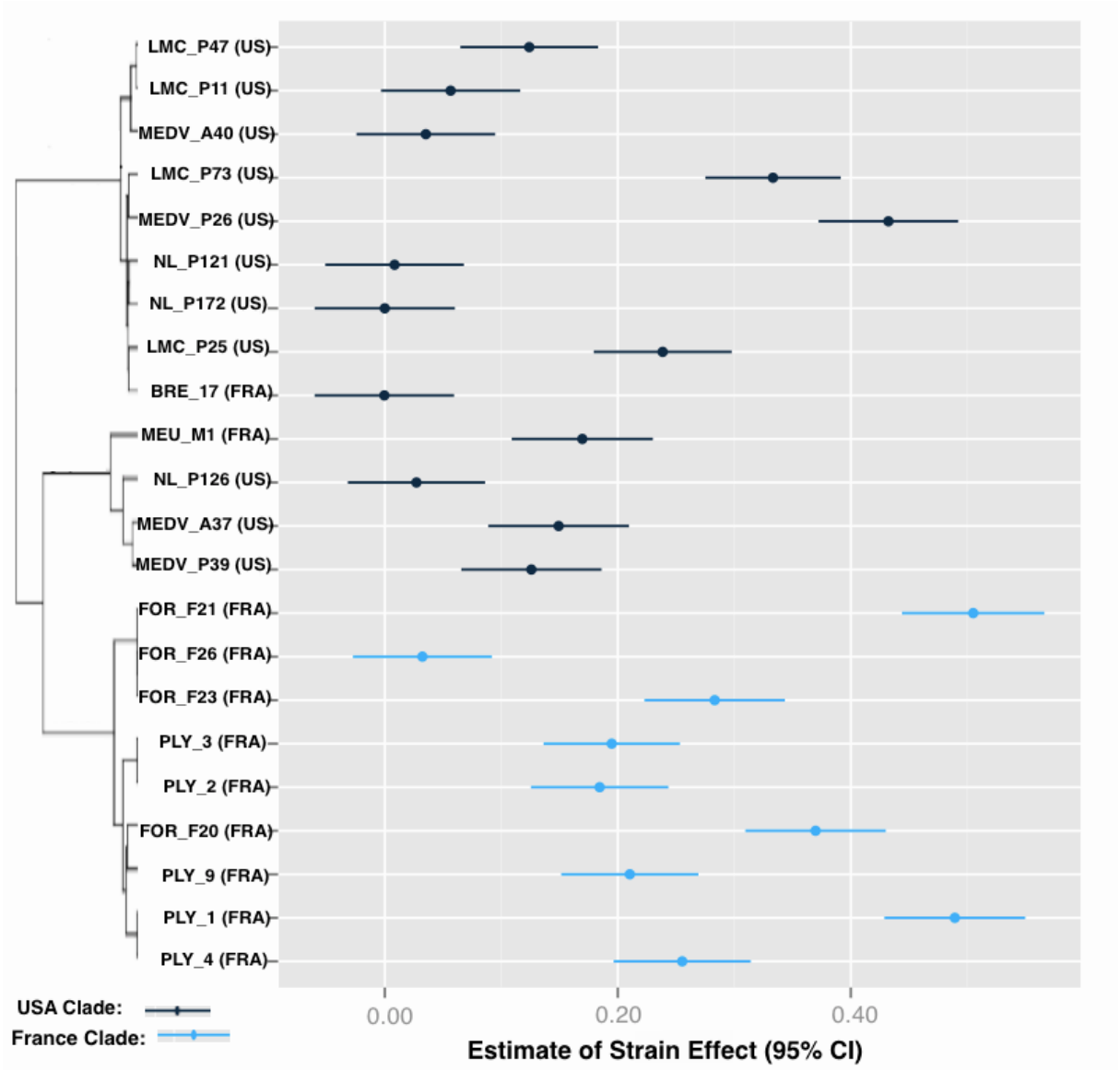

Fig. S3: Marginal effects of the 22 *X. arboricola* strains on QDR. These effects were estimated from a linear model that included indicators for each of the 22 *X. arboricola* strains, the four *A. thaliana* subpopulations, and their interactions as predictors, with correction for additional covariates. The hierarchical clustering overlaid on the left was obtained using UPGMA [17] based on the *X. arboricola* GRM. For the dissimilarity measure we used $d(i,j) = [1 - \hat{\rho}(i,j)]/2$, where $\hat{\rho}(i,j) = \boldsymbol{K}_p(i,j)/\sqrt{\boldsymbol{K}_p(i,i)\boldsymbol{K}_p(j,j)}$ was the estimate of the genomic correlation between strains i and j, for $i,j = 1,\ldots,22$.

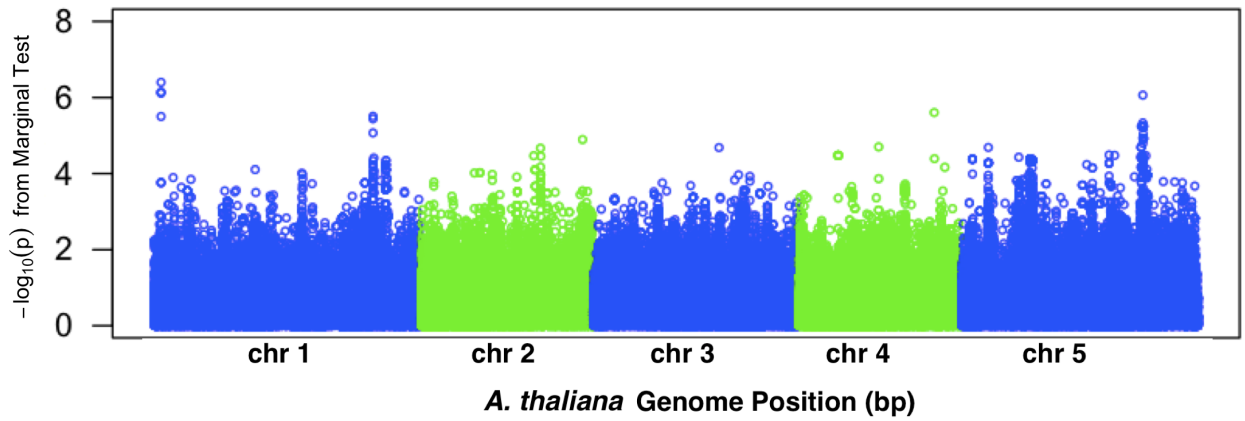

Fig. S4: Manhattan plot of the marginal GWAS for the *A. thaliana* variants. The marginal p-values are plotted against the nucleotide locations (base pairs) on the *A. thaliana* reference genome (version TAIR10). The p-values are obtained from marginal tests under the Gaussian ATOMM model.

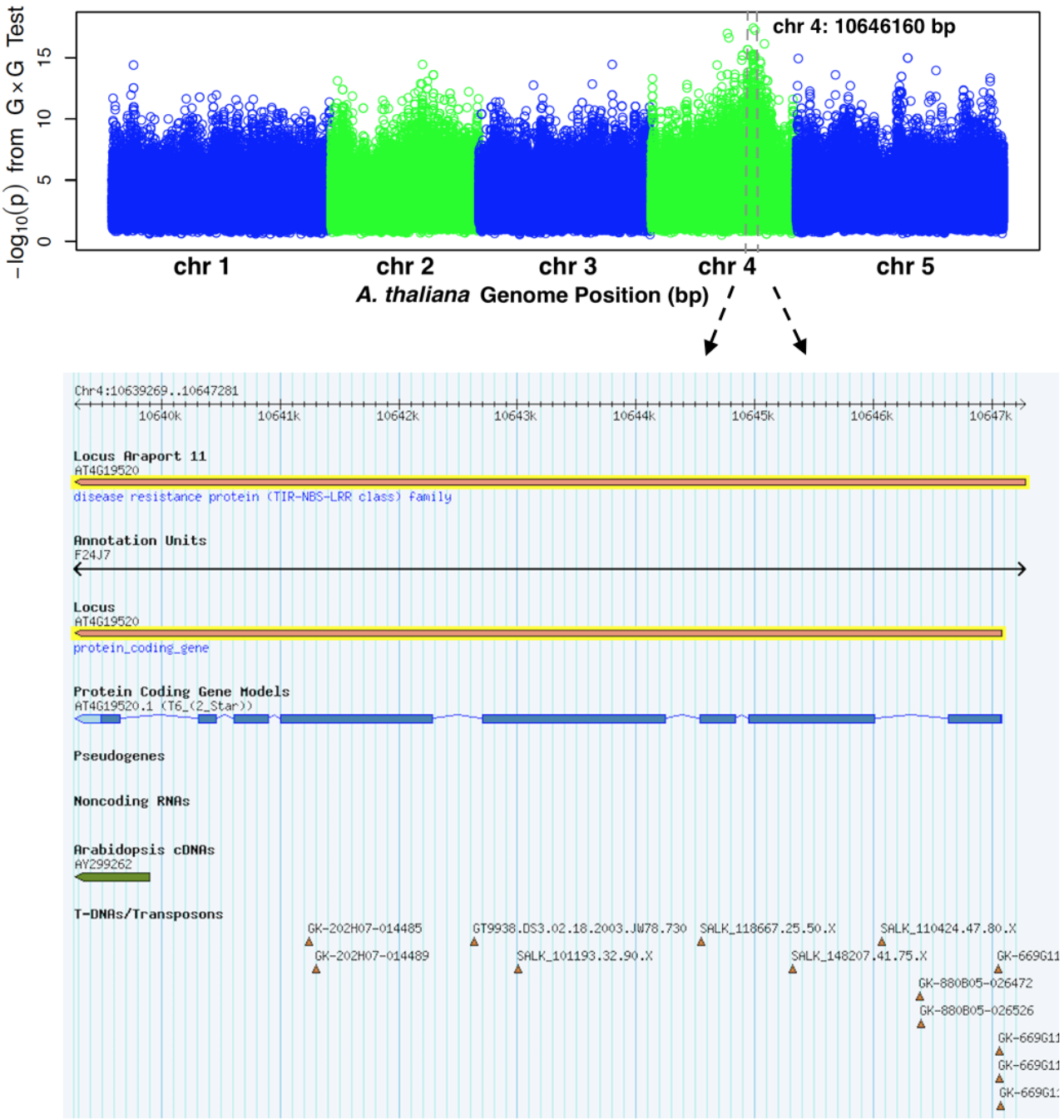

Fig. S5: Genome-wide scan of the *A. thaliana* genome based on G × G tests. The top panel is the Manhattan plot of minimum G × G p-values of the 1,220,413 *A. thaliana* SNPs with MAF ≥ 0.1, where minimum is taken so that each variant will appear only once in the plot. The bottom panel shows the gene annotation model (retrieved from TAIR10) nearby the SNP at location 10646160 bp of chromosome 4. The highlighed SNP (MAF = 0.16) was not among the top marginal effects (marginal p-value = .0757) but was prioritized 2nd in the interaction analysis (G × G p-value = 5.28×10^{-18}). The gene *AT4G19520* is known to encode a disease resistance protein (TIR-NBS-LRR class).

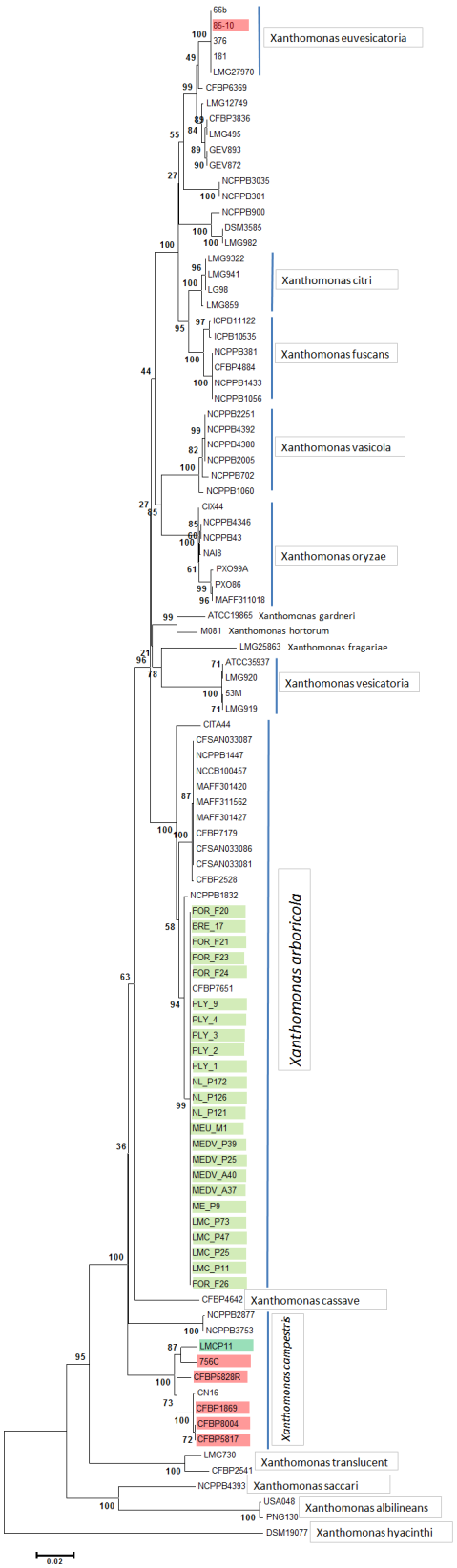

Fig. S6: Phylogeny based on *rpoD* housekeeping gene. Strains labeled in light green were isolated from *A. thaliana*. Strains labeled in red were validated for triggering *RKS1* response in *A. thaliana* [10]. Names of species complex are indicated.

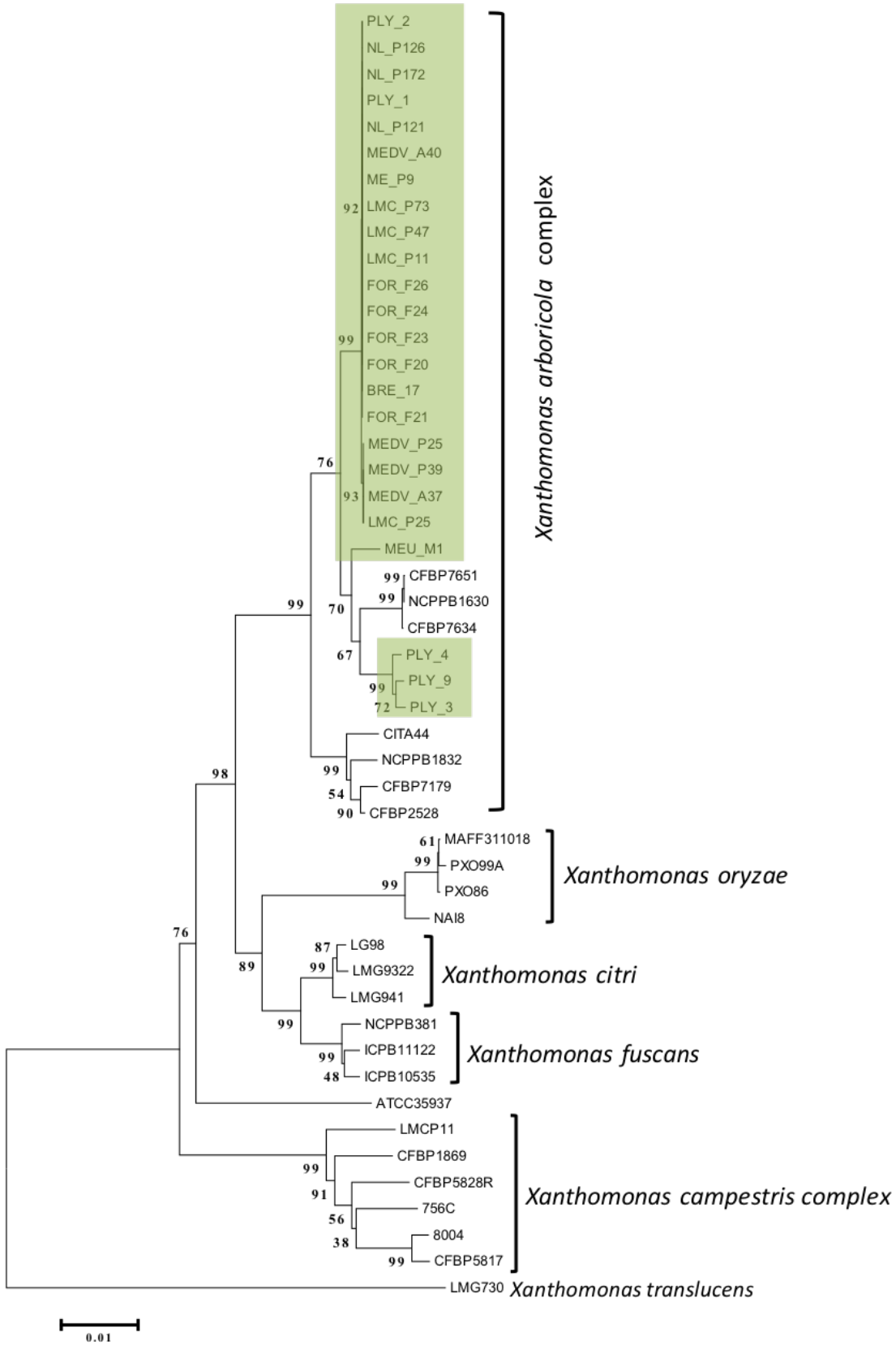

Fig. S7: MLST based on *rpoD*, *gyrB*, *atpD*, and *glnA* housekeeping genes (5865 bp). Strains labeled in green were isolated from *A. thaliana* populations.

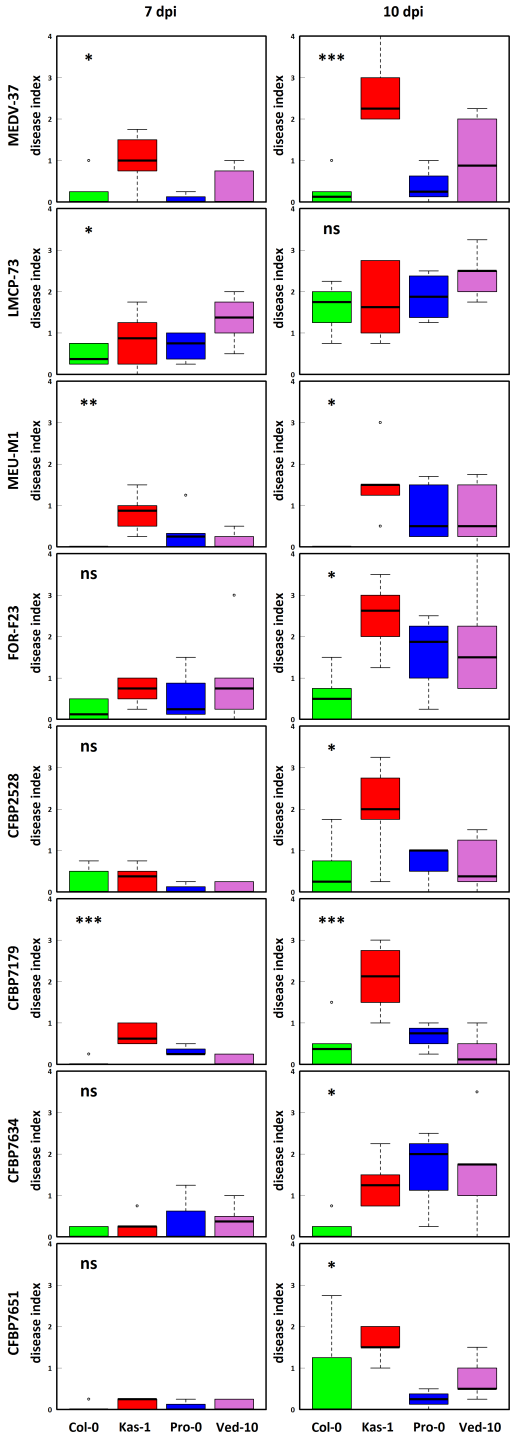

Fig. S8: Pathogenicity of the *X. arboricola* strains isolated from *A. thaliana* vs. crop strains. Pathogenic assays of four strains collected in natural populations of *A. thaliana* (two from USA: MEDV_37 & LMCP_73, two from France: MEU_M1 & FOR_F23) and four crop strains (two reported as pathogenic in the literature: CFBP2528 & CFBP7179, two reported as non-pathogenic in the literature: CFBP7634 & CFBP7651). The x-axis corresponds to four accessions of *A. thaliana* used in this study. '7dpi' and '10dpi' stand for the scoring 7 and 10 days post-inoculation. For each pathogen and each time point, we tested the null hypothesis that the mean QDR was the same for all four accessions of *A. thaliana*. Significance after a Bonferroni correction at a nominal level of 5%: * means P-value between .05 and .001; *** means P-value less than .001.

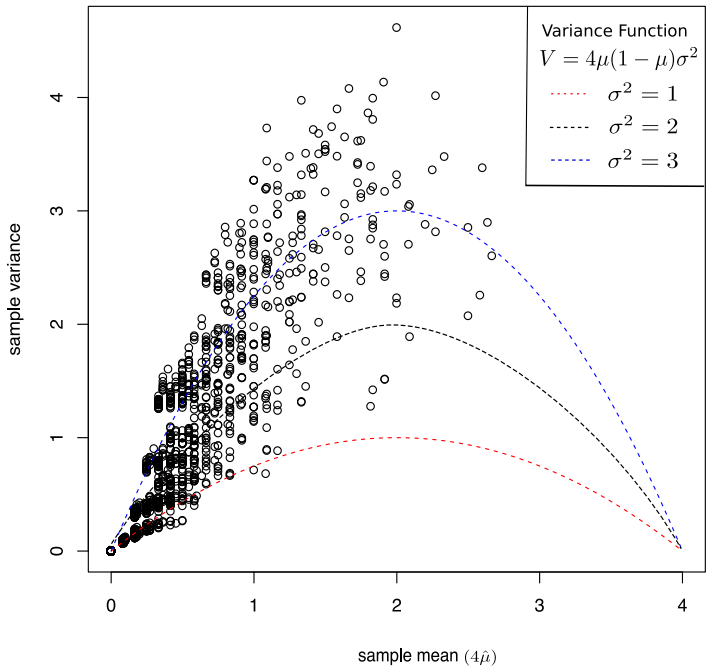

Fig. S9: Empirical relationship between the mean and the variance of QDR. Each point in the figure represents a pairing of an *A. thaliana* line with an *X. arboricola* strain. We hypothesized that each host-strain pair has its own mean and variance. This corresponds to a "full" model where host effects, strain effects, and host-strain interaction effects are included as predictors. We calculated the sample mean and the variance of QDR using replicates for each host-strain pair, while properly taking into account other covariates and block effects (see **Materials and Methods** in the paper). The three curves shown represent three possible values of the dispersion parameter ($\sigma^2 = 1,\ 2,\ 3$).

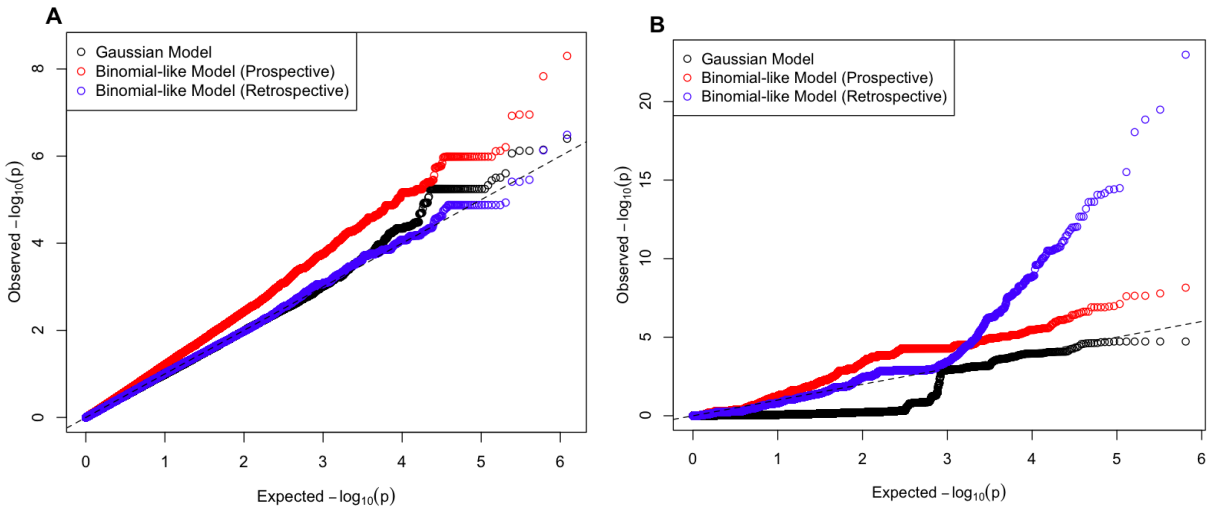

Fig. S10: Q-Q plots of the marginal p-values for the variants on the *A. thaliana* genome (A) and *X. arboricola* genome (B). The observed p-values were obtained under the prospective Gaussian model (black), prospective Binomial-like model (red), and retrospective binomial-like model (blue), respectively. The expected p-values were from a uniform[0,1] distribution. We note that the p-values from the (prospective) Gaussian model are well-calibrated, whereas the p-values from the prospective binomial-like model exhibit modest genome-wide inflation. The p-values from the retrospective binomial-like model appear to be properly calibrated.

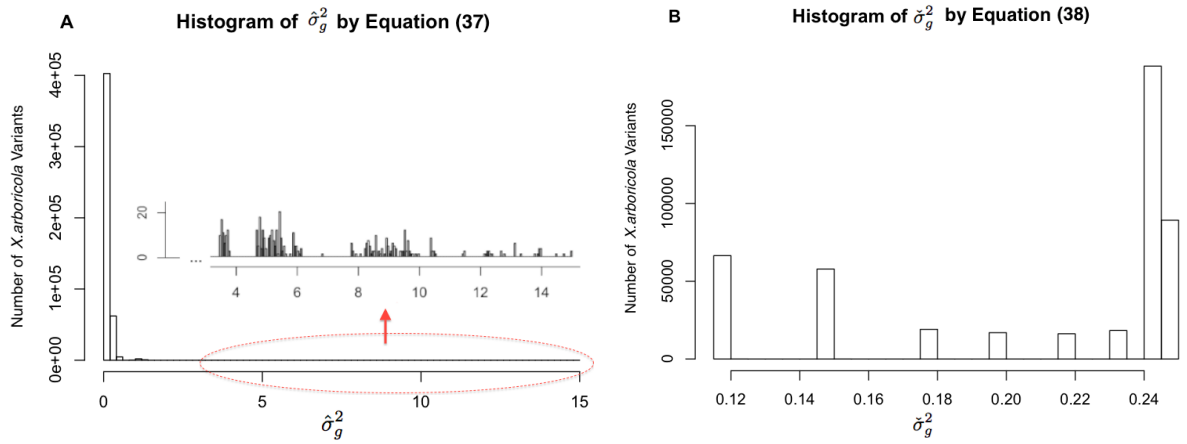

Fig. S11: Comparison of two estimates of σ_g^2 for *X. arboricola* variants. We consider only "core" SNPs, i.e., SNPs that exhibit genotype states $\{0, 1\}$ among the 22 strains in the study. Panel (A) shows the histogram of $\hat{\sigma}_g^2$ based on equation (39) (*SI Appendix, Supplementary Text*), and panel (B) shows the histogram of $\check{\sigma}_g^2$ based on equation (40) (*SI Appendix, Supplementary Text*). Several SNPs exhibited abnormally large values of $\hat{\sigma}_g^2$ in panel (A). We found that these SNPs differentiate the two closely related strain pairs, {FOR_F26, FOR_F21} and {PLY_4, PLY_1}. In contrast, the estimator $\check{\sigma}_g^2$ appears rather stable so we chose to use this estimator in our context.

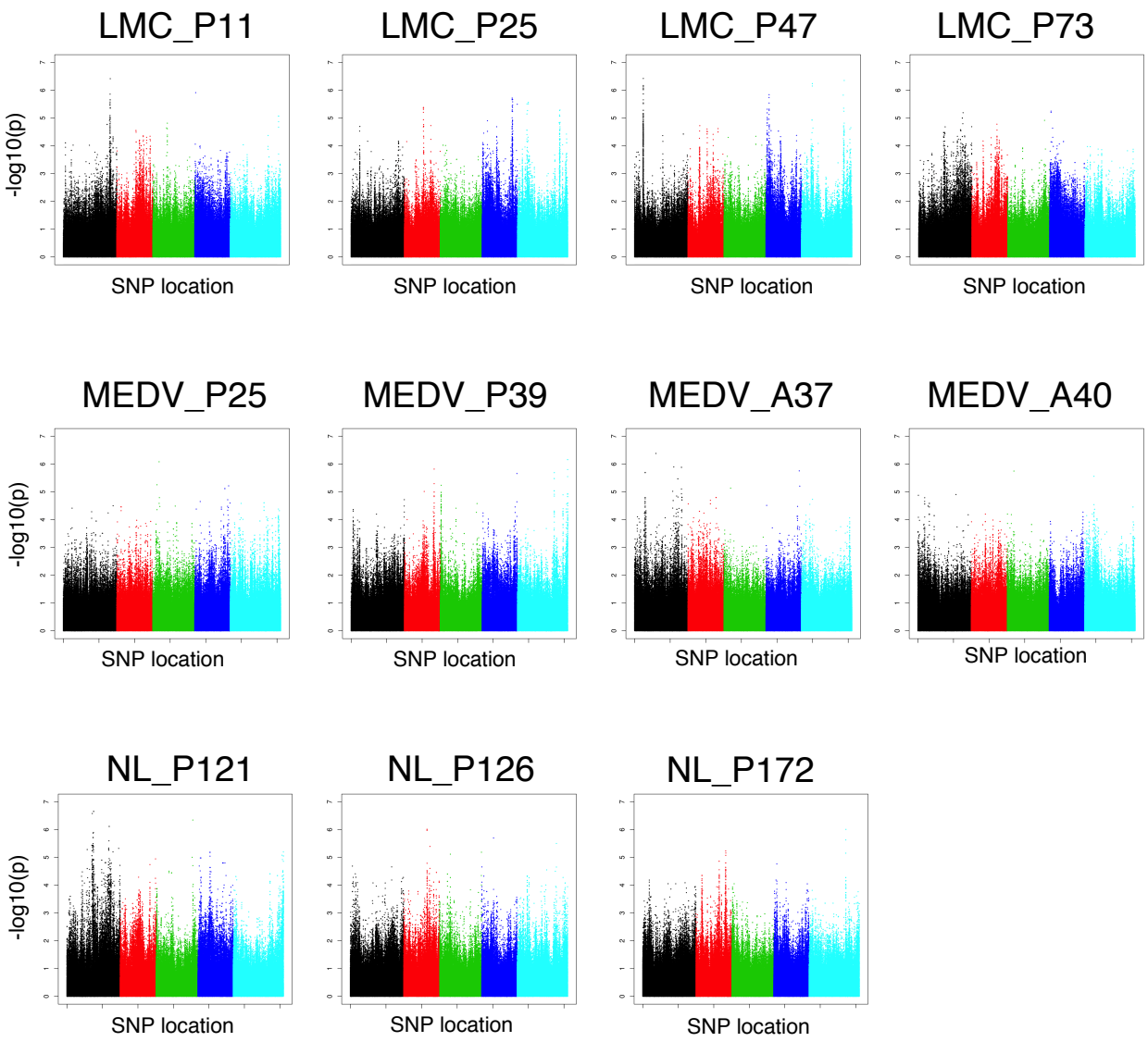

Fig. S12: Separate *Arabidopsis* GWA mapping for US strains. For each of the 11 US strains (LMC_P11, LMC_P25, PMC_P47, LMC_P73, MEDV_P25, MEDV_P39, MEDV_A37, MEDV_A40, NL_P121, NL_P126, NL_P172), we take the subset of the data and then perform genome-wide association analysis for *Arabidopsis* SNPs. The association p-values are plotted against the SNP locations on the *A. thaliana* reference genome. The p-values are obtained from classical linear mixed-effects model.

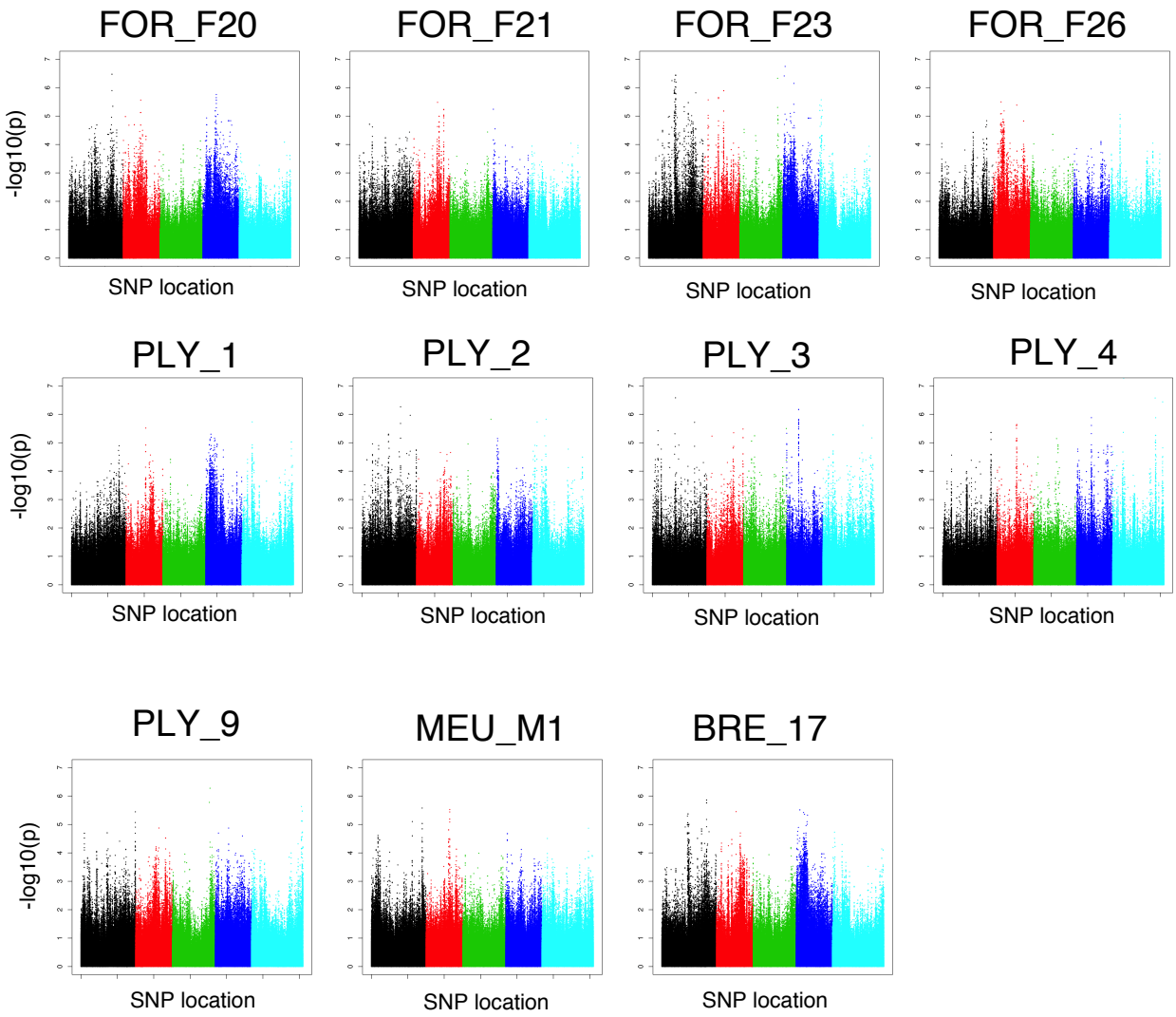

Fig. S13: Separate *Arabidopsis* GWA mapping for France strains. For each of the 11 France strains (FOR_F20, FOR_F21, FOR_F23, FOR_F26, PLY_1, PLY_2, PLY_3, PLY_4, PLY_9, MEU_M1, BRE_17), we take the subset of the data and then perform genome-wide association analysis for *Arabidopsis* SNPs. The association p-values are plotted against the SNP locations on the *A. thaliana* reference genome. The p-values are obtained from classical linear mixed-effects model.

Biological Process	Enrichment	P-value	ATG number	Locus name	Molecular function
regulation of cell division	118.6	**	AT5G51020	CRL - CRUMPLED LEAF	not defined
chloroplast fission	90.3	**	AT5G51020	CRL - CRUMPLED LEAF	not defined
malate metabolic process	68.1	**	AT5G50950	FUMARASE 2	fumarase enzyme
aerobic respiration	64.6	**	AT5G51060	RHD2	NADPH oxidase
regulation of chlorophyll	58.2	**	*AT4G31920*	ARR10	Arabidopsis response regulator (ARR) protein
biosynthetic process			AT5G50920	HSP93-V	protein similar to ATP-dependent ATP-dependent Clp protease ATP-binding subunit / ClpC
response to reactive oxygen species	55.9	**	AT5G51020	CRL - CRUMPLED LEAF	not defined
mitotic recombination	54.9	**	AT5G50930	MHF1	protein with similarity to mammalian MHF1
synapsis	54.8	**	AT5G50930	MHF1	protein with similarity to mammalian MHF1
primary root development	50.9	**	*AT4G31920*	ARR10	Arabidopsis response regulator (ARR) protein
			AT5G51040	SDHAF2	succinate dehydrogenase assembly factor 2 (SDHAF2)
root epidermal cell differentiation	37.2	**	AT5G51060	RHD2	NADPH oxidase
protein targeting to chloroplast	35.3	**	AT5G50920	HSP93-V	protein similar to ATP-dependent Clp protease ATP-binding subunit / ClpC
tricarboxylic acid cycle	28.7	**	AT5G50950	FUMARASE 2	fumarase enzyme
nitrate assimilation	25.4	**	AT5G50950	FUMARASE 2	fumarase enzyme

Table S1: Enrichment of biological process in the 0.01% tail of the top marginal *A. thaliana* SNPs. The significance of enrichment was assessed using a null distribution based on 10,000 permutations from a procedure that takes into account LD patterns. ** means P-value less than 0.01. For each BP term, we also reported which gene corresponding to that term occurred among the top SNP signals to result in the observed enrichment.

	BRE_17	FOR_F20	FOR_F21	FOR_F23	FOR_F24	FOR_F26	LMC_P11	LMC_P25	LMC_P47	LMC_P73	ME_P9	MEDV_A37	MEDV_A40	MEDV_P25	MEDV_P39	MEU_M1	NL_P121	NL_P126	NL_P172	PLY_1	PLY_2	PLY_3	PLY_4	PLY_9	CFBP2528	CFBP7651	CFBP7634	CFBP7179	NCPPB1630	NCPPB1832	CITA44	IVIA2626.1
Hpa1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	1	0	0	0	0	0	0	1	1	0	1	1	1	0	1
Hpa2	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	1	0	0	0	0	0	0	1	1	0	1	1	1	0	1
hpa3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	1
hpa4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
HpaA	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	1	1	0	1	1	1	0	1
hpaB	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	1	0	0	0	0	0	0	1	1	0	1	1	1	0	1
HpaF	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	1
HpaP	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	1	1	0	1	1	1	0	1
HrcC	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	1	0	0	0	0	0	0	1	1	0	1	1	1	0	1
HrcJ	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	1	0	0	0	0	0	0	1	1	0	1	1	1	0	1
HrcN	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	1	0	0	0	0	0	0	1	1	0	1	1	1	0	1
HrcQ	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	1	0	0	0	0	0	0	1	1	0	1	1	1	0	1
HrcR	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	1	0	0	0	0	0	0	1	1	0	1	1	1	0	1
HrcS	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	1	0	0	0	0	0	0	1	1	0	1	1	1	0	1
HrcT	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	1	0	0	0	0	0	0	1	1	0	1	1	1	0	1
HrcU	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	1	0	0	0	0	0	0	1	1	0	1	1	1	0	1
HrcV	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	1	0	0	0	0	0	0	1	1	0	1	1	1	0	1
HrpB1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	1	0	0	0	0	0	0	1	1	0	1	1	1	0	1
HrpB2	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	1	0	0	0	0	0	0	1	1	0	1	1	1	0	1
HrpB3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
HrpB4	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	1	0	0	0	0	0	0	1	1	0	1	1	1	0	1
HrpB5	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	1	0	0	0	0	0	0	1	1	0	1	1	1	0	1
HrpB6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
HrpB7	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	1	1	0	1	1	1	0	1
HrpB8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
hrpD5	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	1	0	0	0	0	0	0	1	1	0	1	1	1	0	1
hrpD6	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	1	1	0	1	1	1	0	1
hrpE1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	1	1	1	0	1
hrpE3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
hrpF1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	1	0	0	0	0	0	0	1	1	0	1	1	1	0	1

Table S2: Composition of the *hrp/hrc* cluster of the 24 studied strains and the 8 crop strains (in dark blue).

	BRE_17	FOR_F20	FOR_F21	FOR_F23	FOR_F24	FOR_F26	LMC_P11	LMC_P25	LMC_P47	LMC_P73	ME_P9	MEDV_A37	MEDV_A40	MEDV_P25	MEDV_P39	MEU_M1	NL_P121	NL_P126	NL_P172	PLY_1	PLY_2	PLY_3	PLY_4	PLY_9	CFBP2528	CFBP7651	CFBP7634	CFBP7179	NCPPB1630	NCPPB1832	CITA44	IVIA 2626.1
AvrBs1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AvrBs2	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	1	0	0	0	0	0	0	1	1	1	1	1	1	0	1
AvrBs3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AvrXccA1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	0	1	0	1	1	1	1	1	1	0	0	0	0	0	0	0	0
AvrXccA2	1	1	1	1	1	1	0	1	0	0	0	0	0	0	0	0	1	0	1	1	1	1	1	1	1	0	0	1	0	0	0	1
HrpW	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	1	0	0	1
XopA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
XopAA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
XopAB	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
XopAC	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
XopAD	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
XopAE	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
XopAF	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
XopAG	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
XopAH	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
XopAI	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
XopAJ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
XopAK	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
XopAL1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
XopAL2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
XopAM	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
XopB	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
XopC1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
XopC2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
XopD	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
XopE1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
XopE2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
XopE3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
XopF1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	1	0	0	0	0	0	0	1	1	0	1	1	1	0	1
XopF2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
XopG	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
XopH	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
XopI	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
XopJ1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
XopJ2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
XopJ3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
XopJ4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
XopJ5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
XopK	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	1
XopL	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	1
XopM	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
XopN	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	1
XopO	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
XopP	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
XopQ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	1	0	0	0	1
XopR	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	0	0	1
XopS	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
XopT	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
XopU	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
XopV	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
XopW	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
XopX	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	1
XopY	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
XopZ1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	1

Table S3: Type III secretion system effector repertoire of the 24 studied strains and the 8 crop strains (in dark blue).

Strain name	Number of proteins	Inparalogs	Inparalogs specific	Specific proteins in single copy
BRE_17	4283	183	0	124
CFBP2528	4282	203	2	43
CFBP7179	4373	202	8	100
CFBP7634	4109	183	0	120
CFBP7651	4191	161	3	100
CITA44	4020	169	0	182
FOR_F20	4328	184	4	175
FOR_F21	4416	181	0	63
FOR_F23	4403	184	0	52
FOR_F24	4388	181	0	48
FOR_F26	4413	186	0	59
IVIA2626	4337	234	12	327
LMC_P11	4255	167	0	109
LMC_P25	4206	184	2	101
LMC_P47	4146	166	0	21
LMC_P73	4107	171	0	87
MEDV_A37	4059	172	2	100
MEDV_A40	4232	153	2	112
MEDV_P25	4199	182	0	158
MEDV_P39	4098	188	4	181
MEU_M1	4265	208	9	272
ME_P9	4173	176	0	57
NCPPB1630	4152	174	0	180
NCPPB1832	4107	141	0	103
NL_P121	4192	195	2	86
NL_P126	3943	163	2	219
NL_P172	4169	192	5	101
PLY_1	4365	192	0	67
PLY_2	4384	212	2	41
PLY_3	4399	204	0	65
PLY_4	4439	232	2	123
PLY_9	4450	202	0	184

Table S4: Statistics for the annotated genomes of the 24 strains in our study (in red) and the eight strains isolated from crops (in black).

	Model		
	Fixed-effects	i.i.d. Random-effects	ATOMM
Proportion of Variance due to	Estimate	Estimate (s.e.)	Estimate (s.e.)
A. thaliana	.21 (.091)	.05 (.006)	.02 (.014)
X. arboricola	.52 (.190)	.03 (.004)	.44 (.121)
A. thaliana–X. arboricola Interaction	.09 (.160)	.05 (.005)	.05 (.019)
Plant	.008 (.006)	.17 (.008)	.09 (.036)
$\hat{\sigma}_t^2$	.78 (.009)	.79 (.008)	1.23 (.010)
Log-likelihood	-36,940	-38,046	-38,037
BIC	103,674	76,144	76,126

Table S5: Parameter estimates under various models for QDR. For the fixed-effects model, the proportion of variance explained by each factor is obtained based on model equation (41), and for $\hat{\sigma}_T^2$, the observed trait variance is used. For the i.i.d. random-effects and ATOMM models, the proportions of variance are estimated by fitting ξ_h, ξ_p, ξ_{hp}, ξ_J and σ_t^2 in model equations (42) and (43), respectively. Note that $\hat{\sigma}_t^2$ for the ATOMM model is not directly comparable to that for the other models because it contains many covariance terms not present in the other models.

	Model		
	Fixed-effects	i.i.d. Random-effects	ATOMM
Proportion of Variance due to:	Estimate	Estimate (s.e.)	Estimate (s.e.)
A. thaliana	.20 (.074)	.06 (.007)	.06 (.019)
X. arboricola	.07 (.030)	.02 (.004)	.03 (.014)
A. thaliana–X. arboricola Interaction	.21 (.057)	.04 (.006)	.18 (.030)
Plant	.11 (.040)	.18 (.009)	.05 (.021)
$\hat{\sigma}_t^2$	.88 (.009)	.87 (.005)	.89 (.010)
Log-likelihood	-27,765	-29,326	-28,783
BIC	75,019	58,702	57,616

Table S6: Parameter estimates in the sub-analysis of *A. thaliana–X. arboricola* data. We excluded from the analysis seven strains with minimal marginal effects (6 US strains and 1 France strain FOR_F26; see *Figure S3*). This retains 22,478 observations from the original 32,960 observations. For the fixed-effects model, the proportion of variance explained by each factor is obtained based on model equation (41), and for $\hat{\sigma}_T^2$, the observed trait variance is used. For the i.i.d. random-effects and ATOMM models, the proportions of variance are estimated by fitting ξ_h, ξ_p, ξ_{hp}, ξ_J and σ_t^2 in model equations (42) and (43), respectively. Note that $\hat{\sigma}_t^2$ for the ATOMM model is not directly comparable to that for the other models because it contains many covariance terms not present in the other models.

Estimate Under the Null:	Model (43)	Model (45)
Parameter	Estimate (s.e.)	Estimate (s.e.)
Intercept (β_0)	.19 (.011)	.26 (.173)
Person 1 (β_1)	.15 (.015)	.16 (.014)
Person 2 (β_2)	.20 (.015)	.19 (.014)
US strain (β_3)	–	-.19 (.140)
Total Residual Variance (σ_t^2)	1.23	1.18
Proportion of Residual Variance due to:		
A. thaliana (ξ_h)	.021	.033
X. arboricola (ξ_p)	.441	.311
A. thaliana–*X. arboricola* Interaction (ξ_{hp})	.048	.034
Plant/Block Effect (ξ_J)	.093	.091
Log-likelihood	-38,037	-38,030
BIC	76,157	76,154

Table S7: Parameter estimates in the *A. thaliana*–*X. arboricola* data. We fitted the two different ATOMM models based on equations (43) and (45), respectively. In model (43), *A. thaliana* line effects, *X. arboricola* strain effects, and their interactions were treated as random effects via the use of GRMs. In model (43), an *X. arboricola* country of origin indicator was included as a fixed effect with the pathogen GRM being modified accordingly. Note that because the pathogen GRM differs, the two models are not nested.