Supplementary Materials

Longitudinal Association Mapping Models Using Di-allelic Markers

Assume that I di-allelic markers $M_j, j = 1, 2, \cdots, I$ are typed in a region of the trait locus Q. For marker M_j, the two alleles are denoted by M_j and m_j with frequencies P_{M_j} and P_{m_j}, respectively (note here the notation M_j can be either marker or allele, whichever applies). Suppose that markers M_j are in Hardy-Weinberg equilibrium (HWE). However, they may be in LD. Denote the measure of LD between trait locus Q and marker M_j by $D_{M_jQ} = P(M_jQ_1) - q_1P_{M_j}$, and the measure of LD between marker M_j and marker M_k by $D_{M_jM_k} = P(M_jM_k) - P_{M_j}P_{M_k}, j, k = 1, 2, \cdots, I$ [Hartl and Clark, 1989; Hedrick, 1987; Lewontin, 1988]. Here $P(M_jQ_1)$ and $P(M_jM_k)$ are frequencies of haplotypes M_jQ_1 and M_jM_k, respectively.

Consider a population sample with N individuals. For the i-th individual, let y_i be his/her quantitative trait value and let G_{ij} be his/her genotype at the marker M_j. A temporal LD regression mixed model extending (2) in the main text at the time t can be defined as

$$y_i(t) = \mu(t) + w_i(t)^\tau \beta(t) + \sum_{j=1}^{I} x_{ij}\alpha_j(t) + \sum_{j=1}^{I} z_{ij}\delta_j(t) + U_i(t) + E_i + \epsilon_i. \tag{S.1}$$

The components of the above model are specified as follows. First, $\mu(t)$ is a non-random overall mean at time t and $\mu(t)$ is unspecified; $w_i(t)$ is a row vector of covariates such as gender, BMI at the time t, and possible their interaction terms; $\beta(t)$ is a non-random column vector of regression parameters of the covariates $w_i(t)$ with fixed effects. One may want to notice that the covariates can be time invariant like gender or can be time varying such as the BMI. In addition, x_{ij} and z_{ij} are dummy variables defined by

$$x_{ij} = \begin{cases} 2 & \text{if } G_{ij} = M_jM_j \\ 1 & \text{if } G_{ij} = M_jm_j \\ 0 & \text{if } G_{ij} = m_jm_j \end{cases} \quad \text{and} \quad z_{ij} = \begin{cases} -P_{m_j}^2 & \text{if } G_{ij} = M_jM_j \\ P_{m_j}P_{M_j} & \text{if } G_{ij} = M_jm_j \\ -P_{M_j}^2 & \text{if } G_{ij} = m_jm_j \end{cases}, \tag{S.2}$$

and $\alpha_j(t)$ and $\delta_j(t)$ are regression coefficients of the dummy variables x_{ij} and z_{ij} at the time t.

In model (S.1), $U_i(t)$ is the correlation effect among repeated measurements of both genetic and environmental factors of an individual, E_i is a random variation of subject i, and ϵ_i is a random measurement error term. Assume that $U_i(t)$, E_i, and ϵ_i are independent. Moreover, assume that E_i is normal $N(0, \sigma_E^2)$ and ϵ_i is normal $N(0, \sigma_e^2)$.

The above formulation is, of course, a general framework. In the context of population data, some of these components may be confounded. In addition, some of the components may need to be removed from the analysis. For instance, assume that the dominance effects are not significantly present. Then model (S.1) can be simplified to

$$y_i(t) = \mu(t) + w_i(t)\beta(t) + \sum_{j=1}^{I} x_{ij}\alpha_j(t) + U_i(t) + E_i + \epsilon_i. \tag{S.3}$$

A similar character process model was developed by Pletcher and Geyer (1999) and Jafferzic and Pletcher (2000), which does not measure effects from specific genes and uses no marker information. The novel part of model (S.1) (or model (S.3)) is that we include measured genotype components estimating association with genotyped markers, i.e., the terms involves x_{ij} and z_{ij} (or x_{ij}) [Fan and Jung, 2003; Fan et al., 2005; Fan and Xiong, 2002; Fan and Xiong, 2003; Jung et al., 2005]. Let the additive and dominance variance-covariance matrices of the indicator variables defined in (S.2) be

$$V_A = 2\begin{pmatrix} P_{M_1}P_{m_1} & D_{M_1M_2} & \cdots & D_{M_1M_I} \\ D_{M_1M_2} & P_{M_2}P_{m_2} & \cdots & D_{M_2M_I} \\ \vdots & \vdots & \cdots & \vdots \\ D_{M_1M_I} & D_{M_2M_I} & \cdots & P_{M_I}P_{m_I} \end{pmatrix}, V_D = \begin{pmatrix} P_{M_1}^2P_{m_1}^2 & D_{M_1M_2}^2 & \cdots & D_{M_1M_I}^2 \\ D_{M_1M_2}^2 & P_{M_2}^2P_{m_2}^2 & \cdots & D_{M_2M_I}^2 \\ \vdots & \vdots & \cdots & \vdots \\ D_{M_1M_I}^2 & D_{M_2M_I}^2 & \cdots & P_{M_I}^2P_{m_I}^2 \end{pmatrix}. \tag{S.4}$$

Such as equation (5) in Jung et al. [2005], the analytical formulas of parameter estimates of models (S.1) and (S.3) at the time t can be obtained as

$$\begin{pmatrix} \alpha_1(t) \\ \vdots \\ \alpha_I(t) \end{pmatrix} = V_A^{-1}\begin{pmatrix} 2D_{M_1Q} \\ \vdots \\ 2D_{M_IQ} \end{pmatrix}\alpha_Q(t) \text{ and } \begin{pmatrix} \delta_1(t) \\ \vdots \\ \delta_I(t) \end{pmatrix} = V_D^{-1}\begin{pmatrix} D_{M_1Q}^2 \\ \vdots \\ D_{M_IQ}^2 \end{pmatrix}\delta_Q(t). \tag{S.5}$$

From equations (S.5), it is clear that the parameters of LD (i.e., D_{M_jQ} and $D_{M_jM_k}$) and gene effects at the time t (i.e., $\alpha_Q(t)$ and $\delta_Q(t)$) are contained in the mean coefficients. Hence, models (S.1) and (S.3) simultaneously take care of the LD and the effects of the putative trait locus Q. Moreover, the interaction between the genetic effects and time or age is modeled.

In the models (S.1) and (S.3), the markers $M_j, j = 1, 2, \cdots, I$, are assumed to be located in a region of a single trait locus Q. This assumption can be removed, i.e., the markers can be from different regions of one chromosome or even from different chromosomes. In one region, there can be one or more trait loci. Thus, the multiple trait loci jointly affect the phenotype. For most interest genetic traits, this is a realistic assumption. Similar arguments as above can be done to justify the models, but notations and formulations can be more complex and we don't provide the details in this article.

Longitudinal Association Mapping Models Using Multi-allelic Markers

In a region of the QTL Q, suppose that multiple multi-allelic markers are typed, which may be micro-satellite markers. For simplicity, we use two marker A and B in our analysis, but the models and methods can be easily generalized to use multiple markers. Suppose that the markers A and B are in HWE. Let us denote the alleles of marker A by $A_1, \cdots, A_a$, where a is the number of alleles. Let the frequency of A_i be $P_{A_i}, i = 1, 2, \cdots, a$. There are $J_A = a(a+1)/2$ possible genotypes, which can be listed as $A_1A_1, \cdots, A_aA_a, A_1A_2, \cdots, A_1A_a, \cdots, A_{a-1}A_a$. The marker B has b alleles denoted by $B_1, \cdots, B_b$. Let the frequency of allele B_k be $P_{B_k}, k = 1, 2, \cdots, b$. There are $J_B = b(b+1)/2$ possible genotypes, which can be listed as $B_1B_1, \cdots, B_bB_b, B_1B_2, \cdots, B_1B_b, \cdots, B_{b-1}B_b$.

Again, consider a population sample with N individuals. For the i-th individual, let y_i be his/her quantitative trait value with genotype G_{Ai} at marker A and genotype G_{Bi} at marker B. Following Fan et al. [2006], consider the following "genotype effect model" under normality

$$\begin{aligned} y_i(t) &= \mu(t) + w_i(t)\beta(t) + \sum_{j=1}^{a-1} x_{Aij}\alpha_{Aj}(t) + \sum_{j=1}^{b-1} x_{Bij}\alpha_{Bj}(t) \\ &\quad + \sum_{1\le j<l\le a} z_{Aijl}\delta_{Ajl}(t) + \sum_{1\le j<l\le b} z_{Bijl}\delta_{Bjl}(t) + U_i(t) + E_i + e_i, \end{aligned} \tag{S.6}$$

where the dummy variables $x_{Aij}, z_{Aijl}, x_{Bij}$ and z_{Bijl} are defined by

$$\begin{aligned} x_{Aij} &= \begin{cases} 2 & \text{if } G_{Ai} = A_jA_j \\ 1 & \text{if } G_{Ai} = A_jA_l, l \neq j \\ 0 & \text{else} \end{cases}, \quad z_{Aijl} = \begin{cases} -P_{A_l}^2 & \text{if } G_{Ai} = A_jA_j \\ P_{A_j}P_{A_l} & \text{if } G_{Ai} = A_jA_l, j \neq l \\ -P_{A_j}^2 & \text{if } G_{Ai} = A_lA_l \\ 0 & \text{else} \end{cases}, \\ x_{Bij} &= \begin{cases} 2 & \text{if } G_{Bi} = B_jB_j \\ 1 & \text{if } G_{Bi} = B_jB_l, l \neq j \\ 0 & \text{else} \end{cases}, \quad z_{Bijl} = \begin{cases} -P_{B_l}^2 & \text{if } G_{Bi} = B_jB_j \\ P_{B_j}P_{B_l} & \text{if } G_{Bi} = B_jB_l, j \neq l \\ -P_{B_j}^2 & \text{if } G_{Bi} = B_lB_l \\ 0 & \text{else} \end{cases}, \end{aligned} \tag{S.7}$$

and $\alpha_{Aj}(t), \alpha_{Bj}(t), \delta_{Ajl}(t), \delta_{Bjl}(t)$ are regression coefficients of the dummy variables at the time t. The other terms of model (S.6) are similar as those of model (S.1). Model (S.6) takes both additive and dominance effects into account [Fan et al., 2006]. Such as model (S.3), model (S.6) can be modified to an "additive effect model" if only the additive effect is modeled, i.e.,

$$y_i(t) \;=\; \mu(t) + w_i(t)\beta(t) + \sum_{j=1}^{a-1} x_{Aij}\alpha_{Aj}(t) + \sum_{j=1}^{b-1} x_{Bij}\alpha_{Bj}(t) + U_i(t) + E_i + e_i. \tag{S.8}$$

In the following, we show that the parameters of LD and gene effects are contained in the regression coefficients. Models (S.6) and (S.8) take care of both the LD and the effects of the trait locus Q. They are valid temporal models to fit association between genetic markers and the trait.

Let $x_{Aij}, x_{Bij}, z_{Aijl}$ and z_{Bijl} be the dummy variables defined by relations (S.7). Denote $X_A = (x_{A11}, \cdots, x_{A1(a-1)})^\tau$, $X_B = (x_{B11}, \cdots, z_{B11(b-1)})^\tau$, and $X_{A\cup B} = (X_A^\tau, X_B^\tau)^\tau$. Let us denote the additive variance-covariance matrix of the dummy variables by $V_A = \mathrm{Cov}(X_{A\cup B}, X_{A\cup B}) = \mathrm{E}\,(X_{A\cup B}X_{A\cup B}^\tau) - \mathrm{E}\, X_{A\cup B}(\mathrm{E}\, X_{A\cup B}^\tau)$. Similarly, let $Z_A = (z_{A112}, \cdots, z_{A11a}, z_{A123}, \cdots, z_{A12a}, \cdots, z_{A1(a-1)a)})^\tau$, $Z_B = (z_{B112}, \cdots, z_{B11b}, z_{B123}, \cdots, z_{B12b}, \cdots, z_{B1(b-1)b)})^\tau$, and $Z_{A\cup B} = (Z_A^\tau, Z_B^\tau)^\tau$. Let us denote the dominance variance-covariance matrix of the indicator variables z_{A1ij}, z_{B1kl} by $V_D = \mathrm{Cov}(Z_{A\cup B}, Z_{A\cup B})$. For $i = 1, 2, \cdots, a$, let us denote $D_{A_iQ} = P(Q_1A_i) - q_1P_{A_i}$, which are measures of LD between QTL Q and marker A. Here $P(Q_1A_i)$ is the frequency of haplotype Q_1A_i. For $k = 1, 2, \cdots, b$, let us denote $D_{B_kQ} = P(Q_1B_k) - q_1P_{B_k}$, which are measures of LD between QTL Q and marker B. Here $P(Q_1B_i)$ is the frequency of haplotype Q_1B_i. For $i = 1, 2, \cdots, a, k = 1, \cdots, b$, let us denote $D_{A_iB_k} = P(A_iB_k) - P_{A_i}P_{B_k}$, which are measures of LD between markers A and B. Here $P(A_iB_k)$ is frequency of haplotype A_iB_k.

Such as Appendix V, Fan et al. [2006], we can show that the regression coefficients of models (S.8) and (S.6) are given by

$$\begin{pmatrix} \alpha_{A1}(t) \\ \vdots \\ \alpha_{A(a-1)}(t) \\ \alpha_{B1}(t) \\ \vdots \\ \alpha_{B(b-1)}(t) \end{pmatrix} \;=\; (V_A/2)^{-1} \begin{pmatrix} D_{A_1Q} \\ \vdots \\ D_{A_{a-1}Q} \\ D_{B_1Q} \\ \vdots \\ D_{B_{b-1}Q} \end{pmatrix} \alpha_Q(t)$$

$$\begin{pmatrix} \delta_{A12}(t) \\ \vdots \\ \delta_{A(a-1)a}(t) \\ \delta_{B12}(t) \\ \vdots \\ \delta_{B(b-1)b}(t) \end{pmatrix} = V_D^{-1} \begin{pmatrix} [P_{A_2}D_{A_1Q} - P_{A_1}D_{A_2Q}]^2 \\ \vdots \\ [P_{A_{a-1}}D_{A_aQ} - P_{A_a}D_{A_{a-1}Q}]^2 \\ [P_{B_2}D_{B_1Q} - P_{B_1}D_{B_2Q}]^2 \\ \vdots \\ [P_{B_{b-1}}D_{B_bQ} - P_{B_b}D_{B_{b-1}Q}]^2 \end{pmatrix} \delta_Q(t). \qquad \text{(S.9)}$$

The elements of matrices V_A and V_D are provided in Appendix V, Fan et al. [2006]. Equations (S.9) show that the parameters of LD (i.e., D_{A_iQ} and D_{B_kQ}) and gene effects (i.e., $\alpha_Q(t)$ and $\delta_Q(t)$) are contained in the regression coefficients. The gene substitution effect $\alpha_Q(t)$ is contained only in α_{Aj}, α_{Bj}; and the dominance effect $\delta_Q(t)$ is contained only in $\delta_{Ajl}(t), \delta_{Bjl}(t)$. Thus, V_A is called additive variance-covariance matrix; and V_D is called dominance variance-covariance matrix. The model (S.6) orthogonally decomposes genetic effect into summation of additive and dominance effects.

Analysis of FHS Data from GAW 13

For the trait of systolic blood pressure, Levy et al. (2000) performed a genome-wide linkage analysis of FHS data and identified a multi-allelic locus GATA25A04 (D17S1299) on chromosome 17 which shows strong linkage with a high LOD score 3.8. We concerned about the association between systolic blood pressure and genotypes of the locus GATA25A04 over people's age. At the locus GATA25A04, there are 8 alleles (184, 188, 192, 196, 200, 204, 208, 212). We investigated the temporal trend of systolic blood pressure related to genetic information, age, and sex. After a thorough model selection by fitting linear mixed model in R [Pinheiro and Bates, 2000], we got a final model as

$$\begin{aligned} y_{ij} &= \mu_0 + t_{ij}\mu_{age} + t_{ij}^2\mu_{age^2} + t_{ij}^3\mu_{age^3} + \text{sex}_i\beta_{sex} \\ &\quad + x_{i,188}\alpha_0 + x_{i,188}t_{ij}\alpha_{age} + x_{i,188}t_{ij}^2\alpha_{age^2} + x_{i,188}t_{ij}^3\alpha_{age^3} + U_i(t_{ij}) + E_i + e_{ij}, \qquad \text{(S.10)} \end{aligned}$$

where $x_{i,188} = \begin{cases} 2 & \text{if } G_i = 188/188 \\ 1 & \text{if } G_i = 188/* \\ 0 & \text{else} \end{cases}$ is the number of allele 188 in the genotype G_i of subject i, and $*$ represents the alleles other than allele 188.

The variance estimations are $\hat{\sigma}_E^2 = 11.21^2$ and $\hat{\sigma}_S^2 = 12.82^2$, and correlation range $\hat{\rho} = 1.84$. The regression results of model (S.10) are presented in Table S.1. In addition, we presented in Table S.1 the regression results of the following model

$$y_{ij} = \mu_0 + t_{ij}\mu_{age} + t_{ij}^2\mu_{age^2} + t_{ij}^3\mu_{age^3} + \text{sex}_i\beta_{sex} + x_{i,188}\alpha_0 + U_i(t_{ij}) + E_i + e_{ij}, \qquad \text{(S.11)}$$

which does not include the time-dependent variables $x_{i,188}t_{ij}$, $x_{i,188}t_{ij}^2$, and $x_{i,188}t_{ij}^3$. From the results of model (S.11) in Table S.1, we can see that allele 188 has no significant effect on the SBP since the high p-value 0.28. However, all three time-dependent variables $x_{i,188}t_{ij}$, $x_{i,188}t_{ij}^2$, and $x_{i,188}t_{ij}^3$ have significant effects on SBP at a significance level 0.05 for model (S.10). Therefore, it is important to include the time-dependent genetic variables in the analysis. Ignoring time trends in genetic effects can make the model invalid.

To understand the temporal trend of SBP, Figure S.1 provides the predicted SBP vs age for male and female, separately. The predicted SBP of male is 4.95 higher than that of female. Interestingly, the allele 188 at locus GATA25A04 almost has no effect on SBP for age interval of (25, 40). However, it does have positive effect for SBP when one's age is older than 40. The allele 188 at locus GATA25A04 can be a risk genetic variant since it may lead to higher SBP for middle aged and old people.

In models (S.10) and (S.11), we do not include the random spline variables u_k and v_k as these of spline models (8) and (9) in the main text. We did fit the models by using spline models (8) and (9), but none of them significantly improves the fitting. Hence, none of them is included in the final model. We also fitted non-parametric linear penalized spline models, but they failed to detect the genetic effect of allele 188 although the random term $\sum_{k=1}^{K} u_k(t_{ij} - \kappa_k)_+$ provides significant result.

Table S.1: Association results of blood systolic pressure and marker GATA25A04 (D17S1299) for Framingham Heart Study Data, Genetic Analysis Workshop 13.

Model	**Coefficient**	**Estimates**	**Std Error**	**t-value**	**P-value**
Model (S.10)	μ_0	130.12519	0.5845056	222.62	< 0.0001
	μ_{age}	0.58538	0.0205328	28.51	< 0.0001
	μ_{age^2}	0.00312	0.0007050	4.43	< 0.0001
	μ_{age^3}	-0.00015	0.0000312	-4.96	< 0.0001
	β_{sex}	-4.95030	0.7489797	-6.61	< 0.0001
	α_0	2.56019	1.4748975	1.74	0.08
	α_{age}	0.15237	0.0665988	2.29	0.02
	α_{age^2}	-0.00418	0.0021415	-1.95	0.05
	α_{age^2}	-0.00023	0.0000944	-2.45	0.01
Model (S.11)	μ_0	130.21348	0.5829787	223.36	< 0.0001
	μ_{age}	0.59855	0.0197430	30.32	< 0.0001
	μ_{age^2}	0.00273	0.0006753	4.04	0.0001
	μ_{age^3}	-0.00018	0.0000297	-5.97	< 0.0001
	β_{sex}	-4.94591	0.7488726	-6.60	< 0.0001
	α_0	1.44832	1.3486853	1.07	0.28

The overall likelihood ratio test of model (S.10) vs. model (S.11) to test $H_0 : \alpha_{age} = \alpha_{age^2} = \alpha_{age^3} = 0$ is 9.09, df = 3, p-value = 0.0281.

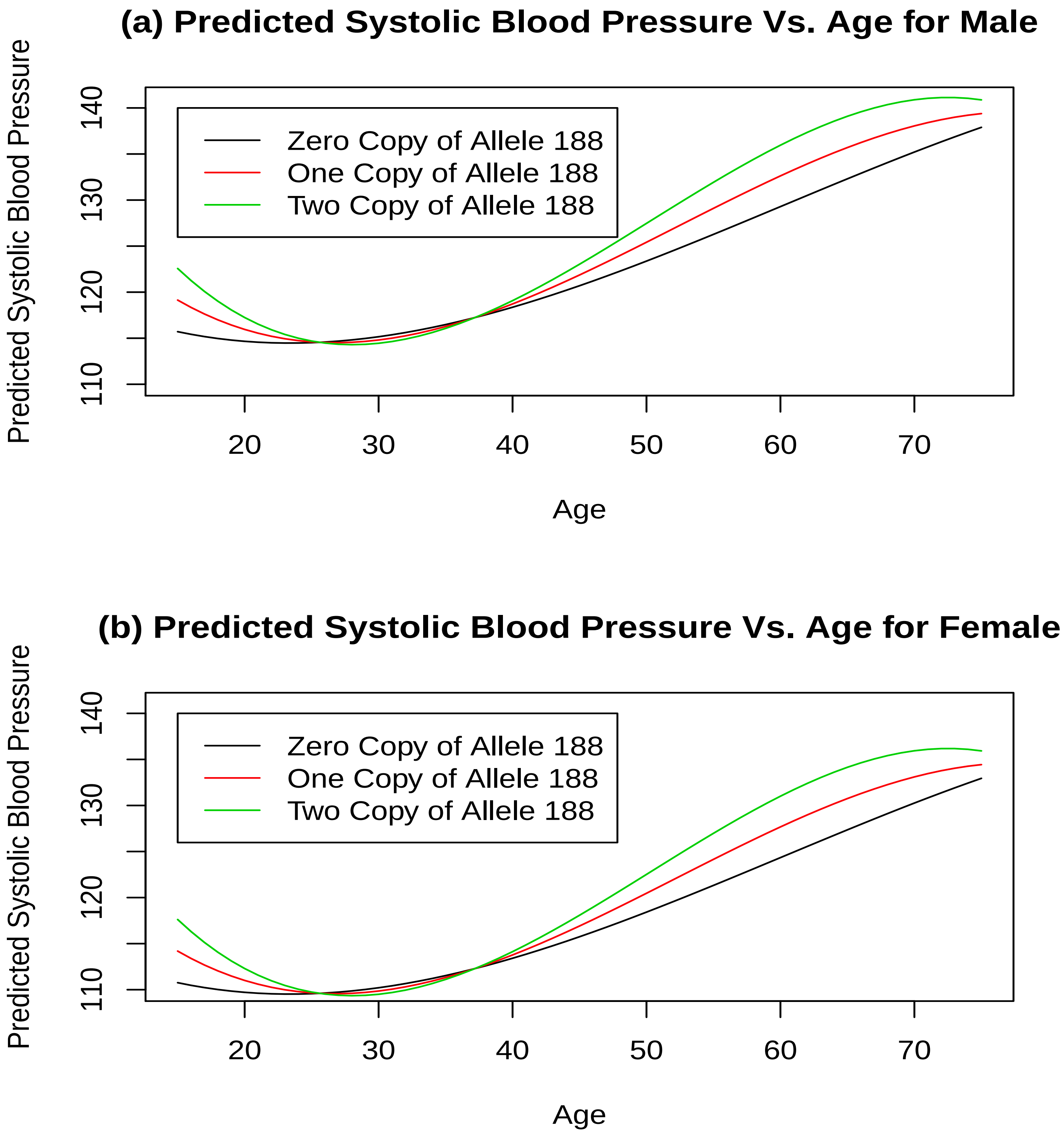

Figure S.1: Predicted systolic blood pressure against age in years for Male and Female by marker GATA25A04 (D17S1299) and sex, based on parametric model (S.10).