Caitlin McHugh
Oct 2014

To investigate association testing on the X chromosome, I implemented simulation studies using various numbers of genotypes that are generated using the pedigree shown in Figure 1.

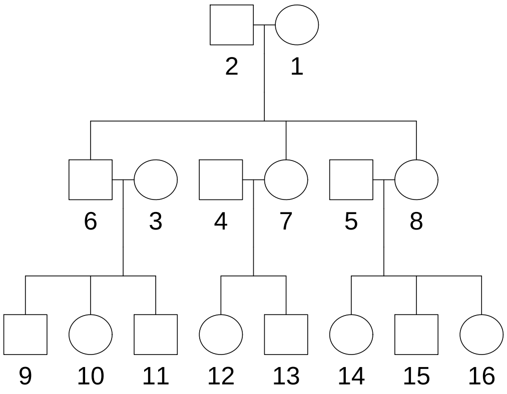

Figure 1: The 16-person pedigree used for the simulations.

The full model assumed when testing for association on X chromosome SNPs (and one of the models used for simulating the X chromosome quantative phenotype) is

$$y = \beta_0 + \beta_1 \text{SNP}_x + g_A + g_X + \epsilon \tag{1}$$
$$g_A \sim MVN(0, \sigma_A^2 \Phi_A) \tag{2}$$
$$g_X \sim MVN(0, \sigma_X^2 \Phi_X) \tag{3}$$
$$\epsilon \sim N(0, \sigma_\epsilon^2) \tag{4}$$

where SNP_x is the vector of genotypes on the X chromosome SNP that is being tested for association, Φ_A is the matrix of kinship coefficients as measured on the autosomes and Φ_X is the matrix of X chromosome specific kinship coefficients. The male X chromosome genotypes are coded as 0, 2 and the female genotypes are coded as 0, 1, 2.

We calculate the variance for a given individual i to be the sum of the variance of the SNP being tested, and the variances due to X chromosome, autosomal and other effects as

$$var(y_i) = \beta_1^2 var(\text{SNP}_x) + \sigma_X^2 + \sigma_A^2 + \sigma_\epsilon^2 \tag{5}$$

The variance of an X chromosome SNP can be calculated conditionally based on whether the sample is female or male.

$$\begin{align}
\mathbb{E}(\text{SNP}_x^F) &= 2p^2 + 2p(1-p) = 2p \tag{6}\\
\mathbb{E}(\text{SNP}_x^M) &= 2p \tag{7}
\end{align}$$

when the male genotypes are coded as 0, 2 and the female genotypes are coded as 0, 1, 2. Then, we find that

$$\begin{align}
var(\text{SNP}_x^F) &= \mathbb{E}((\text{SNP}_x^F)^2) - \mathbb{E}^2(\text{SNP}_x) \tag{8}\\
&= 4p^2 + 2p(1-p) - (2p)^2 \tag{9}\\
&= 2p(1-p) \tag{10}\\
var(\text{SNP}_x^M) &= \mathbb{E}((\text{SNP}_x^M)^2) - \mathbb{E}^2(\text{SNP}_x^M) \tag{11}\\
&= 4p - (2p)^2 \tag{12}\\
&= 4p(1-p) \tag{13}
\end{align}$$

To calculate the covariance of genotypes between a pair of individuals, we must consider their sex. In what follows, I am denoting the X chromosome kinship value between a pair of individuals as Φ_X. Technically, this should be written as $\Phi_{X,ij}$ where ij indexes the individuals i and j for which the X chromosome kinship value represents. First, we calculate the covariance for a SNP between a pair of males as

$$\begin{align}
cov(\text{SNP}_x^M, \text{SNP}_x^M) &= \mathbb{E}(\text{SNP}_x^M\text{SNP}_x^M) - \mathbb{E}^2(\text{SNP}_x^M) \tag{14}\\
&= \mathbb{E}(\text{SNP}_x^M\text{SNP}_x^M|\text{IBD})\mathbb{P}(\text{IBD}) \tag{15}\\
&\ + \mathbb{E}(\text{SNP}_x^M\text{SNP}_x^M|\text{no IBD})\mathbb{P}(\text{no IBD}) - (2p)^2 \tag{16}\\
&= \mathbb{E}(\text{SNP}_x^M\text{SNP}_x^M|\text{IBD})\Phi_X \tag{17}\\
&\ + \mathbb{E}(\text{SNP}_x^M\text{SNP}_x^M|\text{no IBD})(1-\Phi_X) - 4p^2 \tag{18}\\
&= 4p\Phi_X + 4p^2(1-\Phi_X) - 4p^2 \tag{19}\\
&= 4p(1-p)\Phi_X \tag{20}
\end{align}$$

Next we consider the covariance between a pair of female genotypes

$$\begin{align}
cov(\text{SNP}_x^F, \text{SNP}_x^F) &= \mathbb{E}(\text{SNP}_x^F\text{SNP}_x^F) - \mathbb{E}^2(\text{SNP}_x^F) \tag{21}\\
&= \mathbb{E}(\text{SNP}_x^F\text{SNP}_x^F|\text{IBD})\mathbb{P}(\text{IBD}) \tag{22}\\
&\ + \mathbb{E}(\text{SNP}_x^F\text{SNP}_x^F|\text{no IBD})\mathbb{P}(\text{no IBD}) - (2p)^2 \tag{23}\\
&= (2(2p(1-p)) + 4p^2)\Phi_X + 4p^2(1-\Phi_X) - 4p^2 \tag{24}\\
&= 4p(1-p)\Phi_X \tag{25}
\end{align}$$

Finally, we calculate the covariance between a pair of genotypes where one is a female and one is a male

$$
\begin{aligned}
cov(\text{SNP}_x^F, \text{SNP}_x^M) &= \mathbb{E}(\text{SNP}_x^F \text{SNP}_x^M) - \mathbb{E}(\text{SNP}_x^F)\mathbb{E}(\text{SNP}_x^M) && (26)\\
&= \mathbb{E}(\text{SNP}_x^F \text{SNP}_x^M|\text{IBD})\mathbb{P}(\text{IBD}) && (27)\\
&+ \mathbb{E}(\text{SNP}_x^F \text{SNP}_x^M|\text{no IBD})\mathbb{P}(\text{no IBD}) - (2p)^2 && (28)\\
&= [4p^2 + 2(2p(1-p))]\Phi_X + [4p^3 + 4p^2(1-p)](1-\Phi_X) - 4p^2 && (29)\\
&= 4p(1-p)\Phi_X && (30)
\end{aligned}
$$

We can now see that using the X chromosome kinship values from Table 4, the variance for a female and male SNP is indeed as calculated in Equations 10 and 13 after incorporating the self-kinship values. Thus, the variance for a given individual i for an X chromosome SNP is

$$var(y_i) = \beta_1^2 4p(1-p)\Phi_X + \sigma_X^2 + \sigma_A^2 + \sigma_\epsilon^2 \quad (31)$$

The parameter h_{snp}^2 indicates the heritability of the X chromosome SNP. It can be calculated from the equation

$$h_{snp}^2 = \frac{\beta_1^2 4p(1-p)\Phi_X}{\beta_1^2 4p(1-p)\Phi_X + \sigma_\epsilon^2 + \sigma_A^2 + \sigma_X^2} \quad (32)$$

where p is the allele frequency of the causal SNP. On the other hand, we can calculate the heritability of all SNPs on the X chromosome, which is

$$h_x^2 = \frac{\beta_1^2 4p(1-p)\Phi_X + \sigma_X^2}{\beta_1^2 4p(1-p)\Phi_X + \sigma_\epsilon^2 + \sigma_A^2 + \sigma_X^2} \quad (33)$$

Relatedness Estimation Using X Chromosome SNPs

For the proposed model to work, we first must convince ourselves that we can accurately estimate relatedness using genetic material from the X chromosome. Table 4 displays the autosomal and X chromosome kinship coefficients (KC) for a given pair of relatives. The autosomal KC is defined as the probability of sampling two alleles IBD from a given pair of individuals. The X chromosome KC, on the other hand, is defined as the probability of sampling one allele IBD from an X chromosome in a given pair of individuals. Thus, the X chromosome KC will differ from the usual autosomal KC. Furthermore, when calculating the theoretical X chromosome KC, we must take into account the sex of the individuals, whether the pair is related maternally or paternally, and the type of relationship.

X chromosome relatedness can be estimated between two individuals using SNP genotypes. Let j, k, l, m be four individuals such that j, k are male and l, m are female. The genotypes for each of the independent X chromosome SNPs in individual j are denoted by X_{ij} for SNPs $i \in \{1, \ldots, N\}$ and are coded as 0, 1, 2 in females and 0, 2 in males. We can estimate the genetic relatedness (GR) using SNP genotypes, which is twice the X chromsome KC in female-female pairs, $\sqrt{2}$ the X chromosome KC in female-male pairs and equal to the X chromosome KC in male-male pairs. The X chromosome genetic relatedness between two individuals is

$$\text{GR}_{FF} = \frac{1}{N}\sum_{i=1}^{N}\frac{(X_{il}-2p_i)(X_{im}-2p_i)}{2p_i(1-p_i)} \tag{34}$$

$$\text{GR}_{MM} = \frac{1}{N}\sum_{i=1}^{N}\frac{(X_{ij}-p_i)(X_{ik}-p_i)}{p_i(1-p_i)} \tag{35}$$

$$\text{GR}_{MF} = \frac{1}{N}\sum_{i=1}^{N}\frac{(X_{ij}-p_i)(X_{il}-2p_i)}{\sqrt{2}p_i(1-p_i)} \tag{36}$$

To investigate how accurate we can estimate the X chromosome KC, I simulated varying numbers of X chromosome SNPs for one iteration of the 16-sample pedigree shown in Figure 1. The allele frequency of the SNPs was set at 0.4. Figure 2 shows the difference between the estimated and theoretical X chromosome KC for all sample pairs for increasing numbers of SNPs. Figures 3, 4 and 5 show the results broken up by the composition of sex in the related pair. Figure 6 shows a histogram of the estimated X chromosome KC for each relationship type. The true value is shown with a red dotted line and the number of relationships at a given KC value is displayed in the plot title. Many of the relationships have a small sample size. Some relationships are underestimated but all are generally centered around the truth.

As expected, with a larger number of SNPs, we are able to more accurately estimate the true X chromosome KC. We note from all Figures that the estimated KC is at most 0.06 away from the true value. The OLGA genotyping set resulted in approximately 3,500 SNPs on the X chromosome after pruning. Perhaps this number of SNPs, shown in orange in the Figures, should be considered most realistic. We conclude that we are sufficiently able to estimate the X chromosome KC from 3,500 independent, genotyped X chromosome SNPs.

Variance Components Estimation

I estimated the variance components for the autosomes, the X chromosome, and the remaining effects, using the true kinship matrix in both cases. I then fit the mixed model for a quantitative trait on the X chromosome, testing the genotypes simulated on the X chromosome.

Initially, I investigated whether the estimated variance components were converging to what I expected. I simulated 10,000 independent pedigrees as shown in Figure 1 for a total of 16,000 individuals, of whom 5,000 are unrelated (5 founders per pedigree), and performed this simulation 500 times. The relatedness matrices used were the true values, not the estimated ones. I estimated variance components for three models (where the

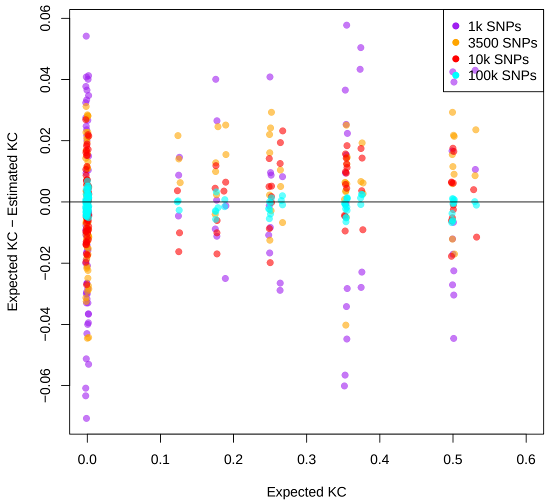

Figure 2: The expected X chromosome KC versus the difference between the expected and estimated X chromosome KC, for all sample pairs in the 16-person pedigree shown in Figure 1. The colors indicate the number of simulated X chromosome SNPs used to estimate the KC.

phenotype was simulated from the given model):

$$y = \beta_1 \text{SNP}_x + \epsilon \tag{37}$$

$$y = \beta_1 \text{SNP}_x + g_X + \epsilon \tag{38}$$

$$y = \beta_1 \text{SNP}_x + g_X + g_A + \epsilon \tag{39}$$

with the specifications of

$$\begin{aligned} g_A &\sim MVN(0, 0.3\Phi_A) \\ g_X &\sim MVN(0, 0.8\Phi_X) \\ \epsilon &\sim N(0,1) \\ \beta_1 &= 0.8 \\ p &= 0.2 \end{aligned}$$

From computations as shown above, in each of the three models we expect the estimates for σ^2_X, σ^2_A and σ^2_ϵ to be as displayed in Table 1 where $\sigma^2_{XT} = \beta_1^2 4p(1-p) + \sigma^2_X$. Table 2 shows the results from the described simulation study. The standard deviations shown there are the mean lower and upper bounds as provided from Matt's confidence intervals, i.e. the confidence intervals calculated from each of the 500 iterations. We note that the simulations yield a mean value that is equal to the expected value. We conclude the variance components are being estimated accurately.

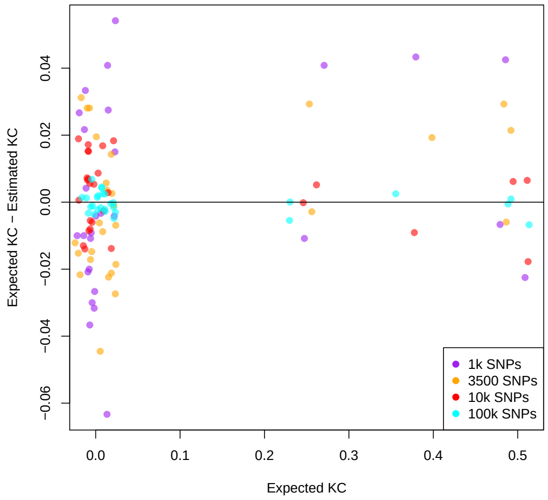

Figure 3: The expected X chromosome KC versus the difference between the expected and estimated X chromosome KC, for all male-male sample pairs in the 16-person pedigree shown in Figure 1. The colors indicate the number of simulated X chromosome SNPs used to estimate the KC.

Association Testing on the X Chromosome

I estimated the variance components for the autosomes, the X chromosome, and the remaining effects, using the known, theoretical autosomal and X chromosome kinship matrices. I fit the mixed model for a quantitative trait on the X chromosome, testing the genotypes simulated on the X chromosome.

We can evaluate the type I error and power when the true model is

$$y = \beta_1 \text{SNP}_x + g_X + g_A + \epsilon \qquad (40)$$

and when fitting the misspecified, usual model, $y = \beta_1\text{SNP}_x + g_A + \epsilon$. I also fit the misspecified model $y = \beta_1\text{SNP}_x + g_X + \epsilon$ for comparison. From 1,000 iterations using 8,000 samples (500 pedigrees as displayed in Figure 1) of which 5,500 are related and 2,500 are unrelated, we set the parameters as shown in Table **??**. Because the usual model is not properly calibrated, to fairly compare the power of each method, we consider a range of α significance levels. For each α value, we calculate the true positive and false positive rate for each model. Then, we can compare the true positive to the false positive rate for each model. In this manner, we can identify how many true positives (the power) we are able to detect for a given false positive rate (type I error).

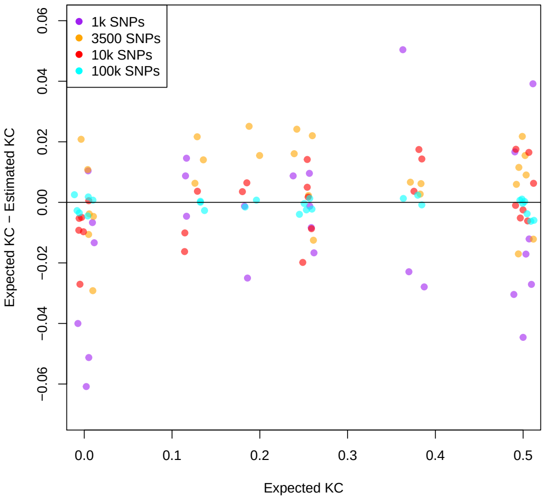

Figure 4: The expected X chromosome KC versus the difference between the expected and estimated X chromosome KC, for all female-female sample pairs in the 16-person pedigree shown in Figure 1. The colors indicate the number of simulated X chromosome SNPs used to estimate the KC.

Performing Mixed Model Association Testing on the X Chromosome

To perform mixed model association testing on X chromsome SNPs, there are a few steps to take.

1. Estimate relatedness on the X chromosome, call the results Φ_X. Use independent (pruned) SNPs, excluding the pseudoautosomal regions. In OLGA, the number of pruned X chromosome SNPs was approximately 3,500. Histograms of the results for theoretical vs estimated in 3,500 simulated X chromosome SNPs for the pedigree in Figure 1 are shown in the file 'hist_xchrKC_byRelType.pdf.'

2. Run Matt's MLM program, including as a random effect the relatedness matrix on the X chromosome, Φ_X. We are investigating the implications of including Φ_A as well, when testing a SNP on the X chromsome for association.

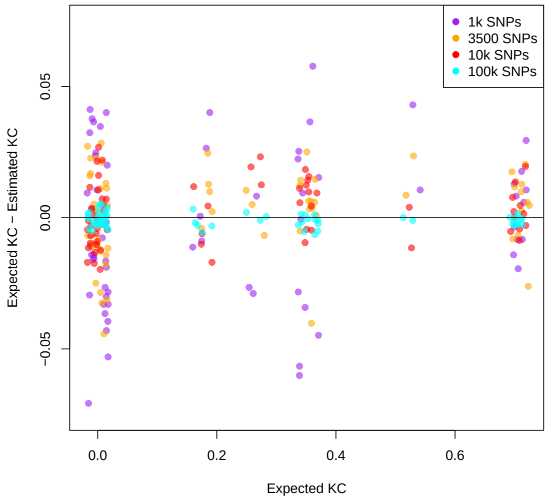

Figure 5: The expected X chromosome KC versus the difference between the expected and estimated X chromosome KC, for all female-male sample pairs in the 16-person pedigree shown in Figure 1. The colors indicate the number of simulated X chromosome SNPs used to estimate the KC.

Model	σ^2_{XT}	σ^2_A	σ^2_ϵ
1	0.4096	-	1
2	1.2096	-	1
3	1.2096	0.3	1

Table 1: Values of simulated variance components.

Model	σ^2_{XT}	σ^2_A	σ^2_ϵ
1	0.4098 (0.3667, 0.4529)	-	1.001 (0.9685, 1.033)
2	1.210 (1.136, 1.284)	-	0.9995 (0.9623, 1.037)
3	1.211 (1.103, 1.319)	0.3035 (0.1303, 0.4768)	1.000 (0.9525, 1.048)

Table 2: Mean (mean CI bounds) from simulation results for three models, where the variance components were simulated as shown in Table 1. The simulation included 16,000 samples, of which 5,000 were unrelated.

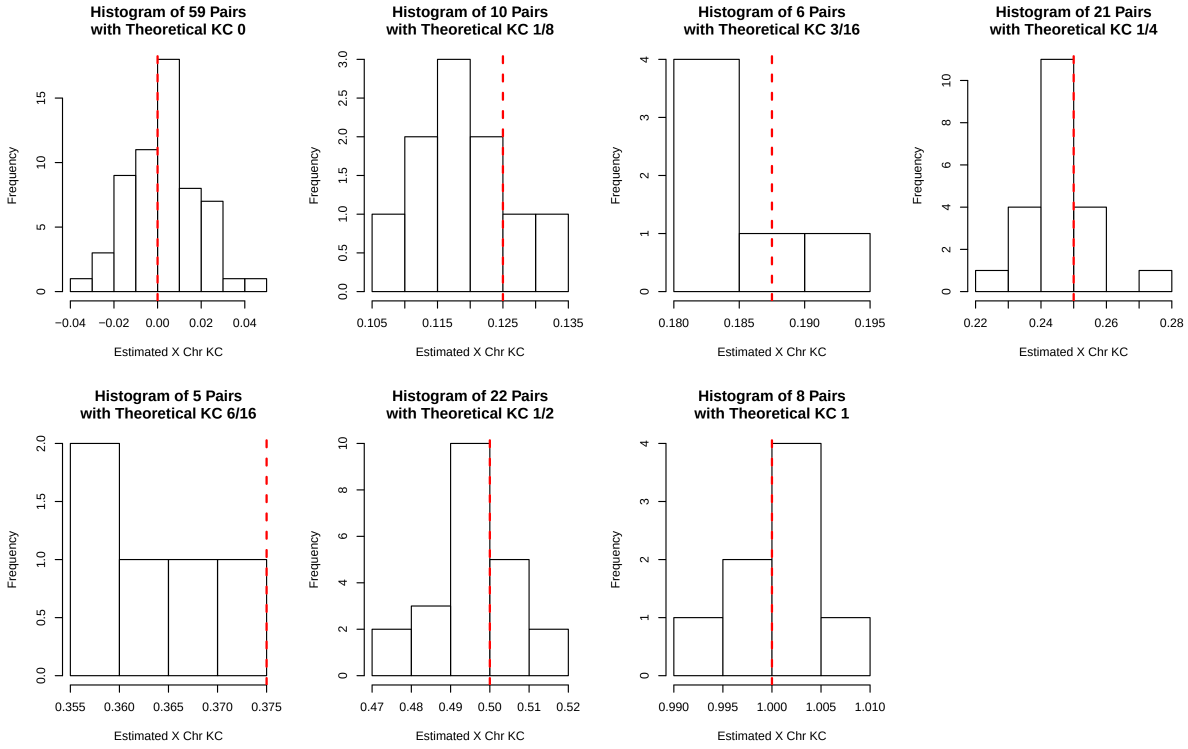

Figure 6: Histograms of the estimated X chromosome KC, by relationship type, for all pairs of relatives shown in the pedigree in Figure 1. The true value is shown with a red dotted line. The estimates were calculated from 3,500 simulated SNPs.

	α	Auto + X (95% CI)	X (95% CI)	Auto (95% CI)
$\sigma^2_X = 0.5, \sigma^2_A = 0.5$	0.05	0.0495 (0.0491, 0.050)	0.0496 (0.0491, 0.050)	0.0587 (0.058, 0.059)
	0.01	0.00987 (0.00968, 0.0101)	0.00984 (0.00964, 0.0100)	0.0131 (0.0129, 0.0133)
	0.005	0.00494 (0.00480, 0.00507)	0.00495 (0.00481, 0.00508)	0.00679 (0.00665, 0.00693)
	0.001	0.00100 (9.39e-04, 0.00106)	0.000988 (9.26e-04, 0.00105)	0.00155 (0.00149, 0.00161)
	5e-04	5.00e-04 (4.56e-04, 5.44e-04)	5.15e-04 (4.71e-04, 5.59e-04)	8.22e-04 (7.78e-04, 8.66e-04)
	1e-04	1.32e-04 (1.12e-04, 1.52e-04)	1.31e-04 (1.11e-04, 1.51e-04)	1.94e-04 (1.74e-04, 2.14e-04)
	5e-05	6.2e-05 (4.81e-05, 7.59e-05)	6.3e-05 (4.91e-05, 7.69e-05)	1.22e-04 (1.08e-04, 1.36e-04)
	1e-05	1.9e-05 (1.28e-05, 2.52e-05)	1.9e-05 (1.28e-05, 2.52e-05)	3.1e-05 (2.48e-05, 3.72e-05)
	5e-06	8e-06 (3.62e-06, 1.24e-05)	5e-06 (6.17e-07, 9.38e-06)	1.8e-05 (1.36e-05, 2.24e-05)
	1e-06	1e-06 (-9.60e-07, 2.96-06)	1e-06 (-9.60e-07, 2.96e-06)	3e-06 (1.04e-06, 4.96e-06)
$\sigma^2_X = 0.5, \sigma^2_A = 0.3$	0.05	0.0499 (0.0495, 0.0503)	0.0499 (0.0494, 0.503)	0.0598 (0.0594, 0.0603)
	0.01	0.0101 (0.00988, 0.0103)	0.0101 (0.00986, 0.0103)	0.0134 (0.0132, 0.0136)
	0.005	0.00504 (0.00490, 0.00518)	0.00504 (0.00490, 0.00517)	0.00711 (0.00697, 0.00725)
	0.001	0.00106 (0.00100, 0.00112)	0.00105 (9.91e-04, 0.00111)	0.00163 (0.00157, 0.00170)
	5e-04	5.31e-04 (4.87e-04, 5.75e-04)	5.32e-04 (4.88e-04, 5.76e-04)	8.69e-04 (8.25e-04, 9.13e-04)
	1e-04	1.01e-04 (8.14e-05, 1.21e-04)	9.8e-05 (7.84e-05, 1.18e-04)	1.92e-04 (1.72e-04, 2.12e-04)
	5e-05	5.0e-05 (3.61e-05, 6.39e-05)	5.1e-05 (3.71e-05, 6.49e-05)	9.7e-04 (8.31e-05, 1.11e-04)
	1e-05	9.0e-06 (2.80e-06, 1.52e-05)	9.0e-06 (2.80e-06, 1.52e-05)	2.1e-05 (1.48e-05, 2.72e-05)
	5e-06	7.0e-06 (2.62e-06, 1.14e-05)	7.0e-06 (2.62e-06, 1.14e-05)	1.1e-05 (6.62e-06, 1.54e-05)
	1e-06	2.0e-06 (4.00e-08, 3.96e-06)	1.0e-06 (-9.60e-07, 2.96e-06)	7.0e-06 (5.04e-06, 8.96e-06)
$\sigma^2_X = 0.3, \sigma^2_A = 0.5$	0.05	0.0500 (0.0495, 0.0504)	0.0500 (0.0496, 0.0504)	0.0530 (0.0526, 0.0535)
	0.01	0.0100 (0.00982, 0.0102)	0.0100 (0.00982, 0.0102)	0.0110 (0.0108, 0.0112)
	0.005	0.00502 (0.00488, 0.00515)	0.00504 (0.00490, 0.00518)	0.00559 (0.00545, 0.00573)
	0.001	9.66e-04 (9.04e-04, 0.00103)	9.79e-04 (9.17e-04, 0.00104)	0.00112 (0.00106, 0.00118)
	5e-04	4.62e-04 (4.18e-04, 5.06e-04)	4.61e-04 (4.17e-04, 5.05e-04)	5.46e-04 (5.02e-04, 5.90e-04)
	1e-04	8.60e-05 (6.64e-05, 1.06e-04)	8.00e-05 (6.04e-05, 9.96e-05)	1.14e-04 (9.44e-05, 1.34e-04)
	5e-05	3.20e-05 (1.81e-05, 4.59e-05)	2.80e-05 (1.41e-05, 4.19e-05)	4.50e-05 (3.11e-05, 5.89e-05)
	1e-05	1.00e-05 (3.80e-06, 1.62e-05)	1.00e-05 (3.80e-06, 1.62e-05)	1.10e-05 (4.80e-06, 1.72e-05)
	5e-06	6.00e-06 (1.62e-06, 1.04e-05)	2.00e-06 (-2.38e-06, 6.38e-06)	8.00e-06 (3.62e-06, 1.24e-05)
	1e-06	0 (-1.96e-06, 1.96e-06)	0 (-1.96e-06, 1.96e-06)	0 (-1.96e-06, 1.96e-06)

Table 3: Type I error from 1,000,000 iterations of an 8,000 sample simulation. The true model is simulated to be $y = \beta_1 \text{SNP}_x + g_A + g_X + \epsilon$ and the results are shown from fitting the true model, the model without fitting X chromosome effects, $y = \beta_1 \text{SNP}_x + g_A + \epsilon$, and the model fitting only X chromosome effects $y = \beta_1 \text{SNP}_x + g_X + \epsilon$.

		Autosomes	X Chromosome
	Self, Female	$\frac{1}{2}$	$\frac{1}{2}$
	Self, Male	$\frac{1}{2}$	1
	Mother-Daughter	$\frac{1}{4}$	$\frac{1}{4}$
	Mother-Son, Father-Daughter	$\frac{1}{4}$	$\frac{1}{2}$
	Father-Son	$\frac{1}{4}$	0
	Full sisters	$\frac{1}{4}$	$\frac{6}{16}$
	Full brothers	$\frac{1}{4}$	$\frac{1}{2}$
	Sister-Brother	$\frac{1}{4}$	$\frac{1}{4}$
Maternal	Aunt-Niece	$\frac{1}{8}$	$\frac{3}{16}$
	Aunt-Nephew	$\frac{1}{8}$	$\frac{6}{16}$
	Uncle-Niece	$\frac{1}{8}$	$\frac{1}{8}$
	Uncle-Nephew	$\frac{1}{8}$	$\frac{1}{4}$
	Grandma-Granddaughter	$\frac{1}{8}$	$\frac{1}{8}$
	Grandma-Grandson	$\frac{1}{8}$	$\frac{1}{4}$
	Grandpa-Granddaughter	$\frac{1}{8}$	$\frac{1}{4}$
	Grandpa-Grandson	$\frac{1}{8}$	$\frac{1}{2}$
Paternal	Aunt-Niece	$\frac{1}{8}$	$\frac{1}{8}$
	Aunt-Nephew	$\frac{1}{8}$	0
	Uncle-Niece	$\frac{1}{8}$	0
	Uncle-Nephew	$\frac{1}{8}$	0
	Grandma-Granddaughter	$\frac{1}{8}$	$\frac{1}{4}$
	Grandma-Grandson	$\frac{1}{8}$	0
	Grandpa-Granddaughter	$\frac{1}{8}$	0
	Grandpa-Grandson	$\frac{1}{8}$	0
Maternal-Maternal	First cousins, Girl-Girl	$\frac{1}{16}$	$\frac{3}{32}$
	First cousins, Girl-Boy	$\frac{1}{16}$	$\frac{3}{16}$
	First cousins, Boy-Boy	$\frac{1}{16}$	$\frac{6}{16}$
Paternal-Paternal	First cousins, Girl-Girl	$\frac{1}{16}$	$\frac{1}{32}$
	First cousins, Girl-Boy	$\frac{1}{16}$	0
	First cousins, Boy-Boy	$\frac{1}{16}$	0
Paternal-Maternal	First cousins, Girl-Girl	$\frac{1}{16}$	$\frac{1}{16}$
	First cousins, Girl-Boy	$\frac{1}{16}$	0
	First cousins, Boy-Boy	$\frac{1}{16}$	0

Table 4: The theoretical kinship coefficients (KC) stratified by X chromosome and autosomes. The autosomal KC value is $\frac{1}{2}\kappa_2 + \frac{1}{4}\kappa_1$, where κ_1 and κ_2 are the probabilities of sampling one and two alleles IBD, respectively. The X chromosome KC value is the probability of sampling one allele IBD on the X chromosome in a given pair of individuals.

		Autosomes	X Chromosome
	Self, Male	$\frac{1}{2}$	1
	Self, Female	$\frac{1}{2}$	$\frac{1}{2}$
	Mother-Son, Father-Daughter	$\frac{1}{4}$	$\frac{1}{2}$
	Full brothers	$\frac{1}{4}$	$\frac{1}{2}$
Maternal	Grandpa-Grandson	$\frac{1}{8}$	$\frac{1}{2}$
	Full sisters	$\frac{1}{4}$	$\frac{6}{16}$
Maternal	Aunt-Nephew	$\frac{1}{8}$	$\frac{6}{16}$
	Mother-Daughter	$\frac{1}{4}$	$\frac{1}{4}$
	Sister-Brother	$\frac{1}{4}$	$\frac{1}{4}$
Maternal	Uncle-Nephew	$\frac{1}{8}$	$\frac{1}{4}$
Maternal	Grandma-Grandson	$\frac{1}{8}$	$\frac{1}{4}$
Maternal	Grandpa-Granddaughter	$\frac{1}{8}$	$\frac{1}{4}$
Paternal	Grandma-Granddaughter	$\frac{1}{8}$	$\frac{1}{4}$
Maternal	Aunt-Niece	$\frac{1}{8}$	$\frac{3}{16}$
Maternal-Maternal	First cousins, Girl-Boy	$\frac{1}{16}$	$\frac{3}{16}$
Maternal-Maternal	First cousins, Boy-Boy	$\frac{1}{16}$	$\frac{6}{16}$
Maternal	Uncle-Niece	$\frac{1}{8}$	$\frac{1}{8}$
Maternal	Grandma-Granddaughter	$\frac{1}{8}$	$\frac{1}{8}$
Paternal	Aunt-Niece	$\frac{1}{8}$	$\frac{1}{8}$
Maternal-Maternal	First cousins, Girl-Girl	$\frac{1}{16}$	$\frac{3}{32}$
Paternal-Maternal	First cousins, Girl-Girl	$\frac{1}{16}$	$\frac{1}{16}$
Paternal-Paternal	First cousins, Girl-Girl	$\frac{1}{16}$	$\frac{1}{32}$
	Father-Son	$\frac{1}{4}$	0
Paternal	Aunt-Nephew	$\frac{1}{8}$	0
Paternal	Uncle-Niece	$\frac{1}{8}$	0
Paternal	Uncle-Nephew	$\frac{1}{8}$	0
Paternal	Grandma-Grandson	$\frac{1}{8}$	0
Paternal	Grandpa-Granddaughter	$\frac{1}{8}$	0
Paternal	Grandpa-Grandson	$\frac{1}{8}$	0
Paternal-Paternal	First cousins, Girl-Boy	$\frac{1}{16}$	0
Paternal-Paternal	First cousins, Boy-Boy	$\frac{1}{16}$	0
Paternal-Maternal	First cousins, Girl-Boy	$\frac{1}{16}$	0
Paternal-Maternal	First cousins, Boy-Boy	$\frac{1}{16}$	0

Table 5: The theoretical kinship coefficients (KC) stratified by X chromosome and autosomes. The autosomal KC value is $\frac{1}{2}\kappa_2 + \frac{1}{4}\kappa_1$, where κ_1 and κ_2 are the probabilities of sampling one and two alleles IBD, respectively. The X chromosome KC value is the probability of sampling one allele IBD on the X chromosome in a given pair of individuals.

Association Testing on the X Chromosome: An Application to HCHS/SOL

Red Blood Cell Traits

I performed association tests on the HCHS/SOL genotyping data, examining a blood cell count trait, red blood cell LABA2 (`316987`). There have been previously published hits on the X chromosome that replicated in SOL. Thus, it was a good candidate for testing out some methodology. To perform the final association test, I fit five models including or excluding fixed and random effects for X chromosome structure specifically. There were a total of 12,488 samples included (see presentation from trait for exclusion criteria).

Need to discuss exercise of estimating x chr KC after adjustment for either autosomal or x chr pcs. note that x KC adj for auto pc still yields structure, implying that pop structure is different on x and something remains beyond what is detected on the autosomes. we want to include the autosomal pcs for precision, even though we don't think that the autosomal population structure directly affects the x chromosome genotypes, ie. we aren't concerned about confounding due to autosomal ancestry.

To calculate the X chromosome principal components (PCs), we first find a pruned set of X chromosome SNPs. To do this, we took all SNPs on the X chromosome that passed the quality filter, of which there were 50,781. Then, for all 10,272 unrelated study samples (unrelated calculated from the autosomes, at degree 4) excluding PC outliers (Asian and Central American outliers), we pruned to a set of independent SNPs using LD pruning with a threshold of 0.32 and a sliding window size of 10Mb. This resulted in exactly 3,600 independent SNPs on the X chromosome. To calculate the PCs, I used `pcair` on the pruned set of X chromosome SNPs. A set of 10,287 unrelated samples was specified (autosomal unrelated at degree 4, excluding outliers as before) and PCs were projected onto 2,497 relatives for a total of 12,784 samples. The first two EVs showed structure, while the EVs 3 and higher did not identify any large-scale structure among the samples. Thus, EVs 1-2 will be used as fixed effects in the MLM-X model.

The X chromosome KC was calculated using `pcrelate` on the LD-pruned set of 3,600 X chromosome SNPs. Again, an unrelated set of 10,287 samples was specified. A total of 12,771 scans was included in the KC matrix, which is 12,784 as above with a further 13 scans excluded due to anomalies that spanned the entire X chromosome. The KC matrix was ancestry-adjusted for X chromosome EVs 1-2 as described above. We thus have a KC matrix that estimates relatedness on the X chromosome after adjusting for X chromosome population structure captured in the first two PCs on the X chromosome.

The estimated proportion variance from each of the components included in the model is as displayed in Table 6. Items to note:

- prop variance explained by autosomal polygenic effects doesn't change with the addition of x chromosome, rather, the environment proportion goes down
- x chromosome polygenic effects explain ∼3%, whereas the autosomal chromosomes 1-22 explain ∼28%, or ∼1.3% per chromosome. the X explains more than is explained by a single autosome, assuming each autosome contributes equally

- x chromosome polygenic effects are non-zero for this trait
- x chromosome effects are non-zero, even when including the autosomal polygenic effects. indicates autosomal effects can't explain all genetic effects
- the var components show an x chromosome effect for this trait, even without an association test

	autosomal KC	block group	environment	household	X chromosome KC
auto PC 1-5 auto KC	0.28473	0.00363	0.66218	0.04945	
X PC 1-2, auto PC 1-5 X KC, auto KC	0.28453	0.00396	0.63266	0.04950	0.02935
X PC 1-2, auto PC 1-5 X KC		0.00271	0.87855	0.08710	0.03164
X PC 1-2 X KC, auto KC	0.28457	0.00402	0.63103	0.05043	0.02995
X PC 1-2 X KC		0.00273	0.87692	0.08806	0.03228

Table 6: Estimated proportion variance for each of the components.

The λ values estimated from fitting these models on X chromosome SNPs specifically, and across the genome (autosomal + X chromsome SNPs) resulted as shown in Table 7. The λ values could be high due to true association, but the same SNPs are considered for all rows of this table. Also, the heightened λ value on the autosomal model indicates just using autosomal structure/relatedness doesn't capture all the structure seen on the x chromosome. Also interesting that highest λ values for X chr, genome-wide, are those adjusted for the wrong type of structure/relatedness. this was shown here with the λ values, ie. trait specific, but also non-trait related in the pc adj for autosomal vs x chromosome structure.

BMI

I performed association tests on the HCHS/SOL genotyping data, examining log(BMI) (298888). In the stratified analysis examining the samples from Central America, there appeared to be a peak of X chromosome SNPs. We hoped to further explore that hit using MLM-X. To perform the final association test, I fit five models including or

	X chromosome	genome-wide
auto PC 1-5 auto KC	1.116	1.051
X PC 1-2, auto PC 1-5 X KC, auto KC	1.043	1.048
X PC 1-2, auto PC 1-5 X KC	1.063	1.106
X PC 1-2 X KC, auto KC	1.049	1.053
X PC 1-2 X KC	1.067	1.109

Table 7: The λ genomic inflation factor calculated from association testing on the X chromosome with the RBC trait, using observed SNPs and filtering using the composite filter and MAC>30, MAF> 0.0012 for a total of 43,868 SNPs. The same filters were used for the genome-wide set of SNPs for a total of 2,128,491 SNPs across the autosomes and X chromosome.

excluding fixed and random effects for X chromosome structure specifically. There were a total of 1,370 samples included (see presentation from trait for exclusion criteria).

To calculate X chromosome PCs to include for the stratified analyses, we first found a pruned set of X chromosome SNPs using all study samples (see subsection above). Then, the subest of unrelated.deg4 samples that fall into each gengrp6.strata category were specified as the unrelated samples, and the remaining gengrp6.strata samples were the set to include for the X chromosome PCA. At the end, we have six new sets of X chromosome EVs, the first two of each we will use as fixed effects in the association tests. We used exactly the same X chromosome KC as described for the red blood cell traits above.

The estimated proportion variance explained from each of the components included in the five models is displayed in Table 9. The genome-wide inflation factor, λ, is presented in Table 10 for all five models.

	Effect Size (SE)	Stat	p-value
autosomal	0.1314 (0.0145)	81.9377	1.40447e-19
X PC 1-2, auto PC 1-5 X KC, auto KC	0.1300 (0.0148)	76.8729	1.82321e-18
X PC 1-2, auto PC 1-5 X KC	0.1318 (0.0149)	78.4084	8.37946e-19
X PC 1-2 X KC, auto KC	0.1301 (0.0148)	77.1621	1.57487e-18
X PC 1-2 X KC	0.1315 (0.0149)	78.3093	8.81049e-19

Table 8: P-value, effect size and standard error for the X chromosome index SNP in all five models. The SNP is Hg19 position 153,764,217 (Xq28) with a MAF in our sample of 12,489 of 0.01979, C/T (T minor allele) and rsID kgp30626516. This position corresponds to dbSNP rsID 1050828, exactly as previously reported by the COGENT study.

	autosomal KC	block group	environment	household	X chromosome KC
autosomal	0.43880	0	0.40434	0.15686	
X PC 1-2, auto PC 1-5 X KC, auto KC	0.42959	0	0.42027	0.15014	0
X PC 1-2, auto PC 1-5 X KC		0	0.78487	0.21513	0
X PC 1-2 X KC, auto KC	0.43545	0	0.41583	0.14872	0
X PC 1-2 X KC		0	0.78727	0.21273	0

Table 9: Estimated proportion variance for each of the components.

	X chromosome	genome-wide
auto PC 1-5 auto KC	1.045	
X PC 1-2, auto PC 1-5 X KC, auto KC	1.000	
X PC 1-2, auto PC 1-5 X KC	1.032	
X PC 1-2 X KC, auto KC	0.9964	
X PC 1-2 X KC	1.035	

Table 10: The λ genomic inflation factor calculated from association testing on the X chromosome, using observed SNPs and filtering using the composite filter and MAC>30, MAF> 0.0111 for a total of 34,424 SNPs. The same filters were used for the genome-wide set of SNPs for a total of 2,128,491 SNPs across the autosomes and X chromosome.

	bkgrd	CentralAm	SouthAm	Mexican	Cuban	Other/Unk	PR	Dominican
	n	1094	686	3635	1732	326	1703	897
CAnD	America	1.049e-64	3.850e-45	1.094e-225	1.043e-53	4.593e-14	4.457e-97	1.868e-18
	Europe	4.883e-70	1.064e-48	2.425e-236	1.563e-77	4.086e-21	6.413e-77	6.178e-46
	Africa	5.489e-01	8.646e-01	3.590e-01	1.705e-24	1.653e-05	5.224e-03	9.273e-23
t-test	America	1.910e-28	2.787e-13	2.605e-100	4.076e-35	3.104e-04	1.041e-80	2.3653e-14
	Europe	3.381e-35	3.758e-16	5.741e-111	2.106e-25	5.660e-07	7.665e-47	1.583e-24
	Africa	7.311e-01	9.305e-01	4.282e-01	8.762e-07	5.877e-02	5.956e-02	1.747e-10

Table 11: Results from applying the CAnD test to the HCHS/SOL subjects, stratified by self-identified background.