Association Testing with X Chromosome Data
An Application To HCHS/SOL

Caitlin McHugh, with Tim Thornton

Department of Biostatistics
University of Washington

5 Feb 2015

Outline

- For each genotyped *autosomal* SNP, we can fit a model of the form

$$Y = \beta_0 + \beta_1 \text{SNP} + g_A + \text{covariates} + \epsilon$$

where

$$g_A \sim MVN(0, \sigma_A^2 \mathbf{\Phi_A})$$

$$\epsilon \sim MVN(0, \sigma_\epsilon^2 \mathbb{I})$$

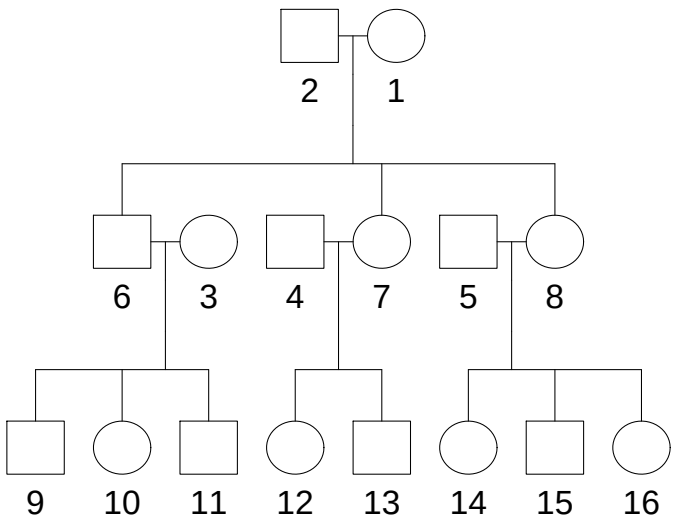
- When testing association on X chromosome SNPs, we propose to fit the model

$$Y = \beta_0 + \beta_1 \text{SNP}_X + g_A + g_X + \text{covariates} + \epsilon$$

where further

$$g_X \sim MVN(0, \sigma_X^2 \mathbf{\Phi_X})$$

- The X chromosome kinship coefficient between individuals i and j, Φ_{ij}^{X}, is defined as the probability of sampling one allele IBD at random from individual i and individual j on the X chromosome.
- *Note* for males there is no randomness in sampling, as there is only one allele at each location on the X chromosome.



		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

- We can estimate $\mathbf{\Phi_X}$ using the following GRM equations:

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

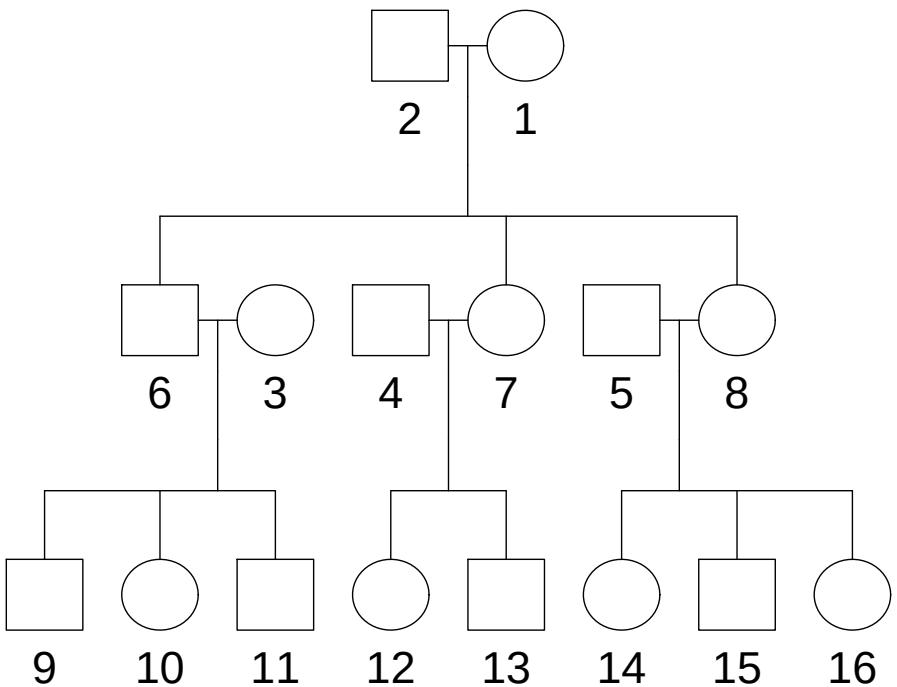
$$\mathrm{GR}_{MM} = \frac{1}{N} \frac{\sum_{i=1}^{N} (X_{ij} - p_i)(X_{ik} - p_i)}{\sum_{i=1}^{N} p_i(1 - p_i)}$$

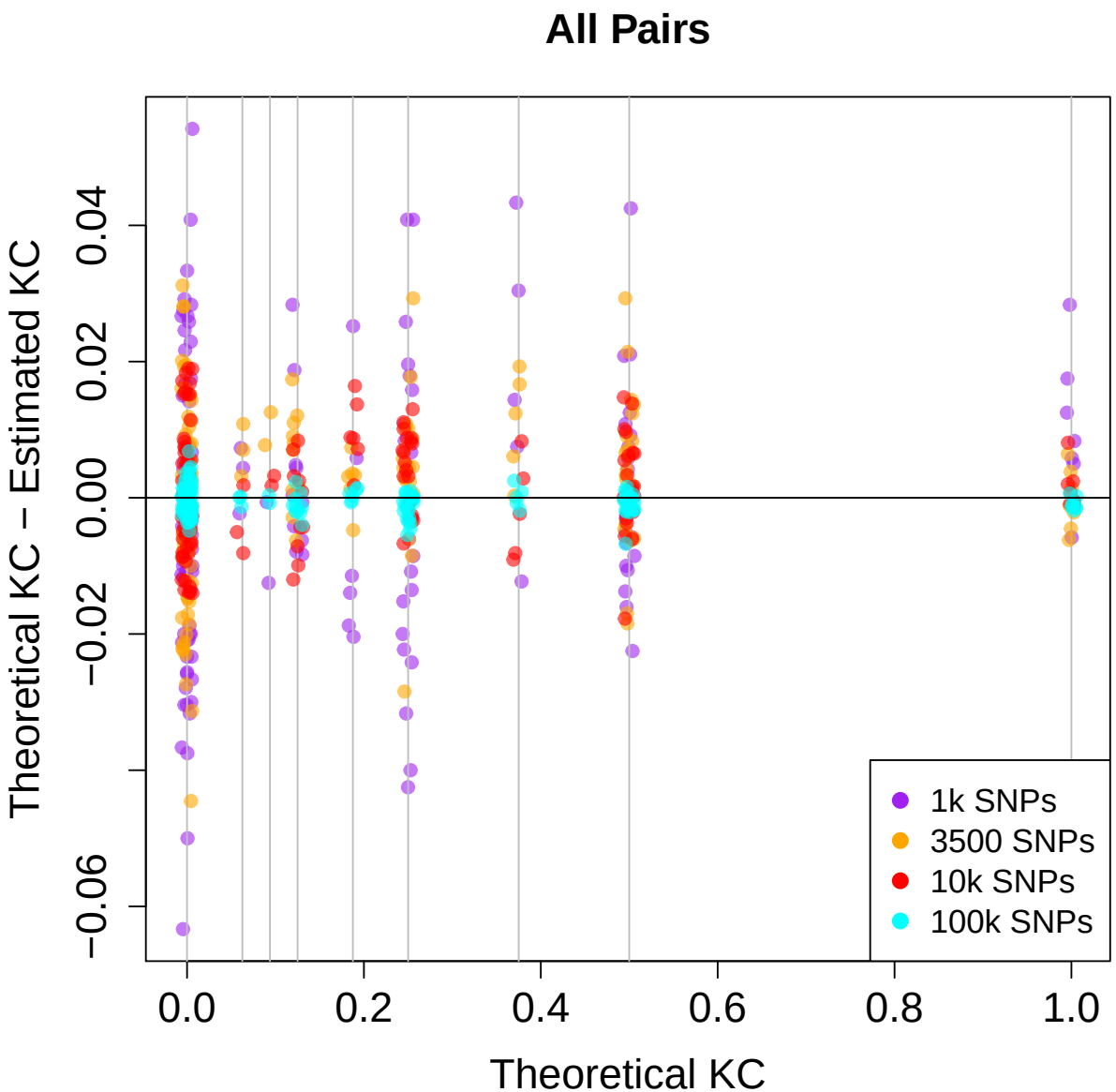
$$\mathrm{GR}_{MF} = \frac{1}{N} \frac{\sum_{i=1}^{N} (X_{ij} - p_i)(X_{il} - 2p_i)}{\sum_{i=1}^{N} \sqrt{2} p_i(1 - p_i)}$$

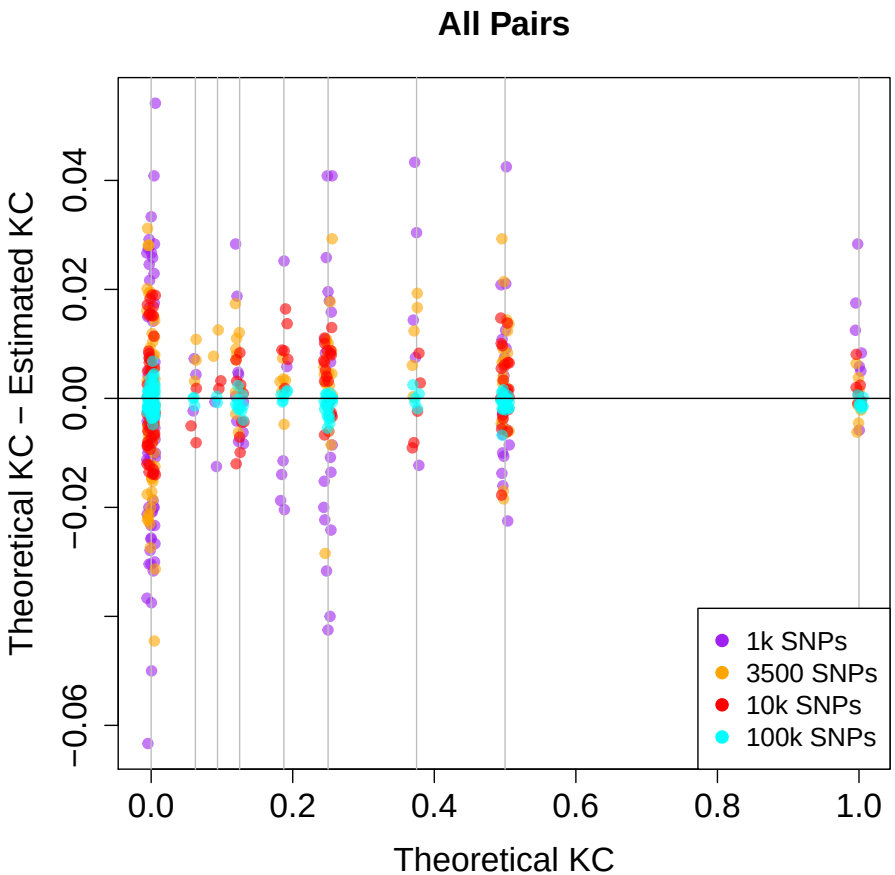
- When X chromosome genotypes are coded 0, 1, 2 for females and 0, 2 for males, the GRM equations for the X chromosome are the same for the autosomes.
- With this genotype coding, the covariance between two individuals i and j on the X chromosome is $4p(1-p)\Phi^X_{ij}$, regardless of the sex of the individuals. {with proof, if needed}

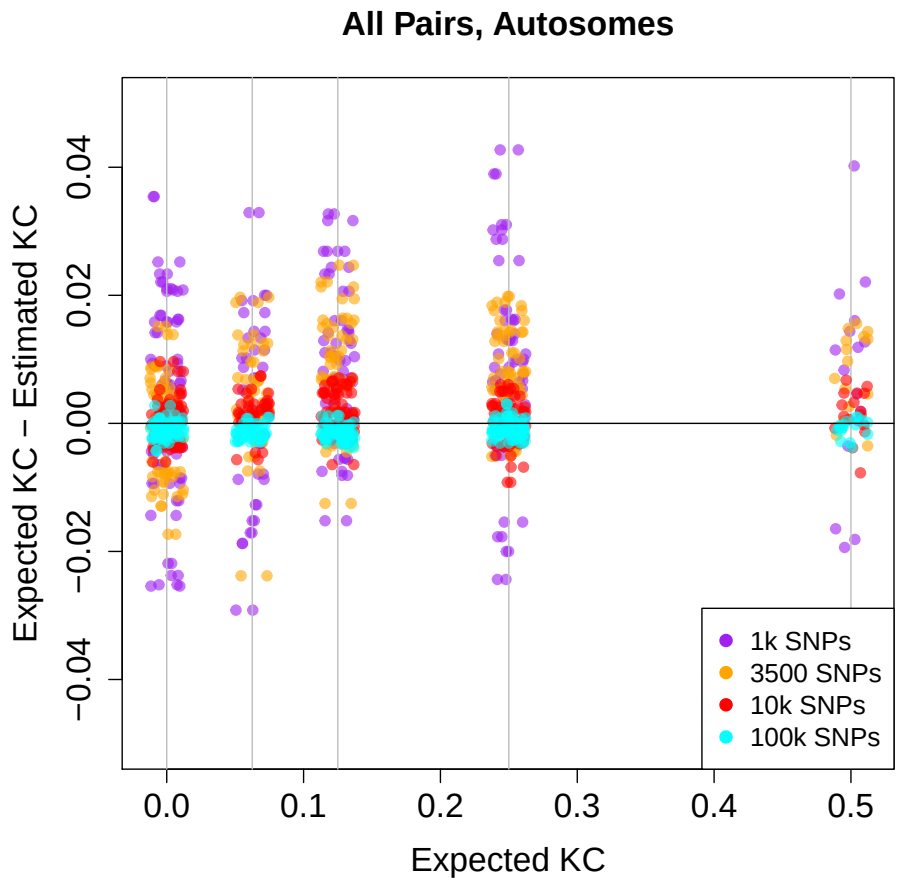
- Can we accurately estimate relatedness using X chromosome SNPs?
- After pruning, there are approximately 3,500 X chromosome SNPs in the SoL data.

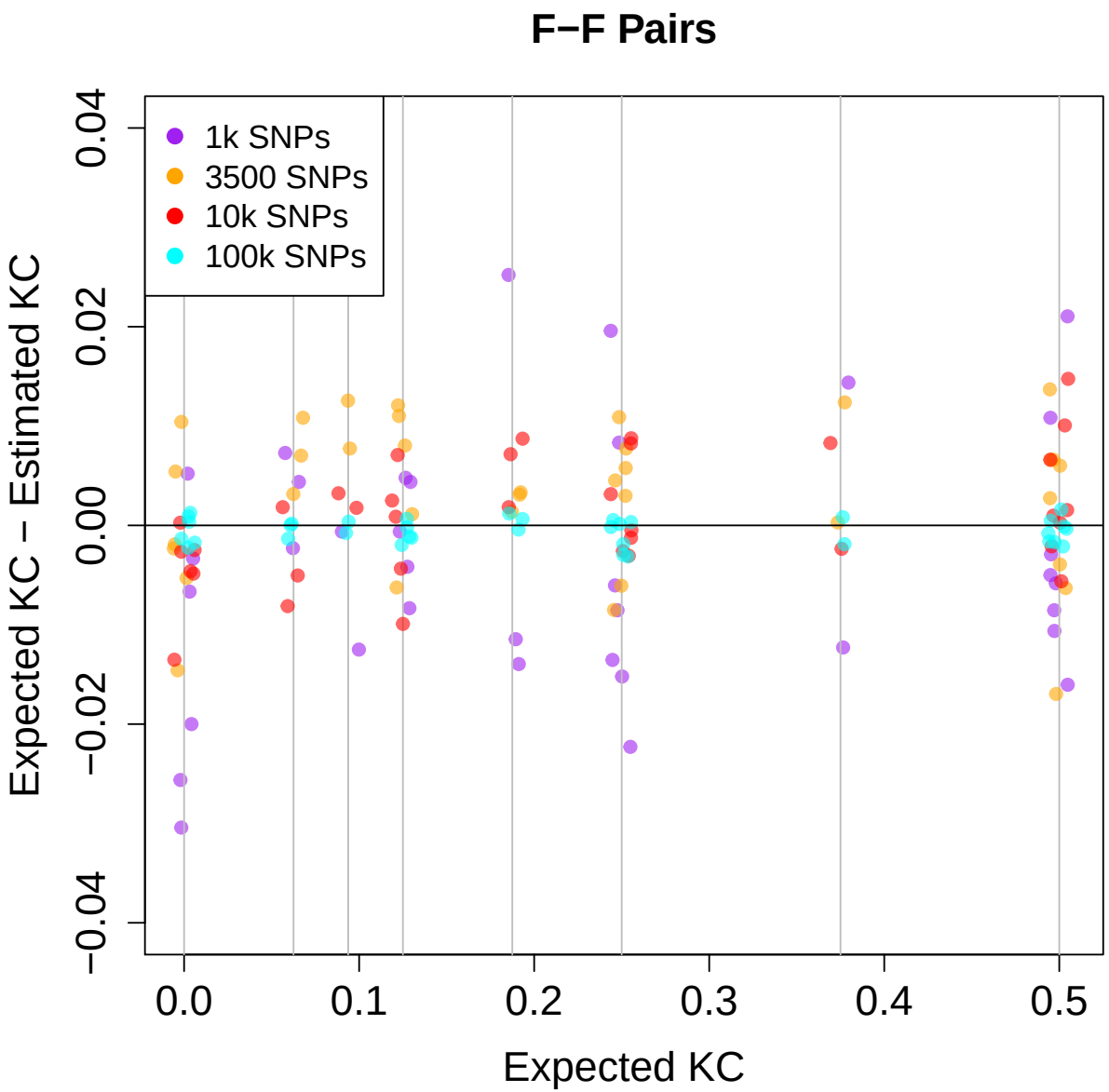
- Simulate increasing number of X chromosome SNPs for 16 individuals in this pedigree structure.
- *Note* this is assuming a homogeneous population.
- Individs 2 and 6 have $\Phi^X_{2,6} = 0$ but $\Phi^A_{2,6} = 1/4$.

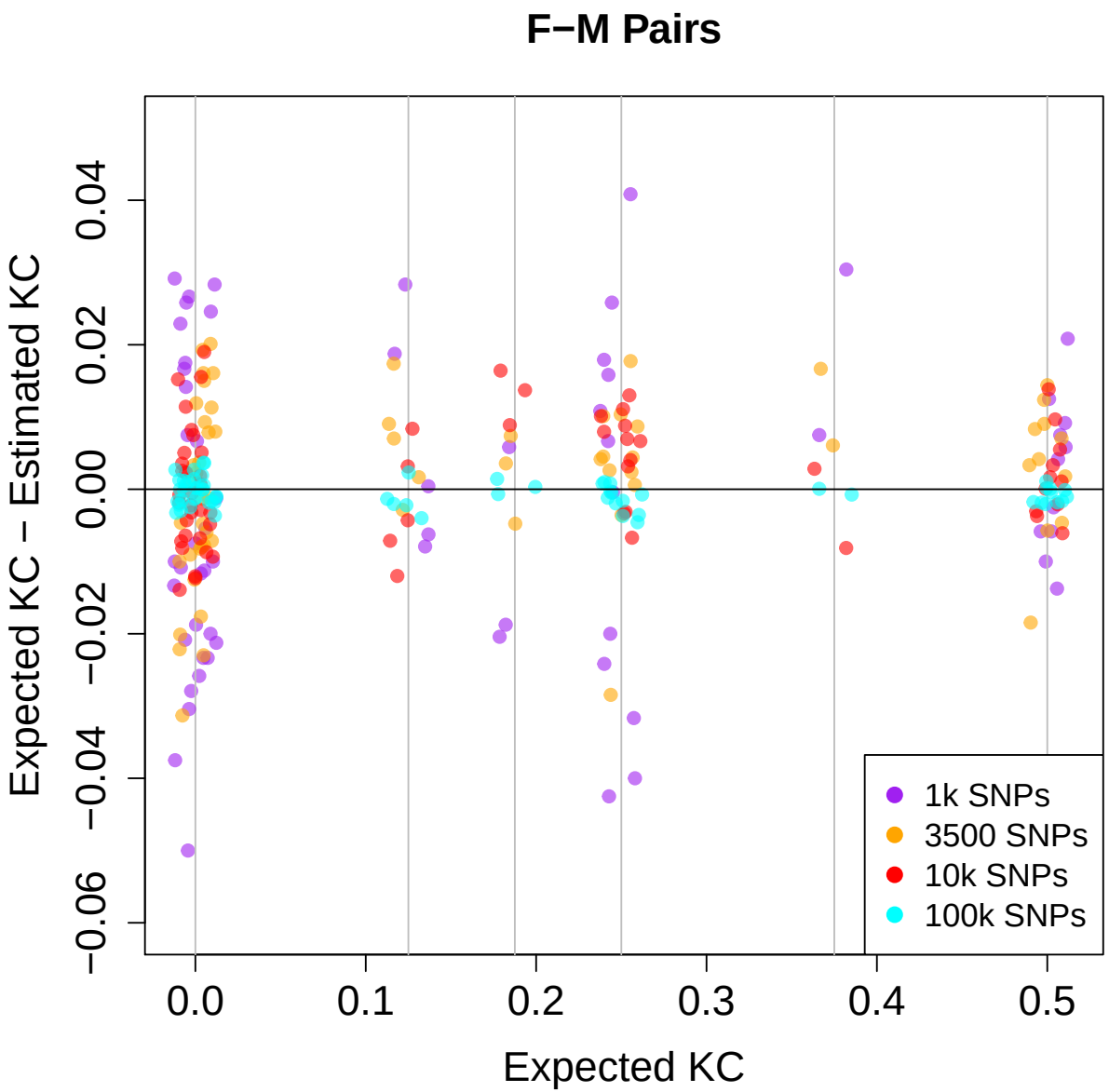


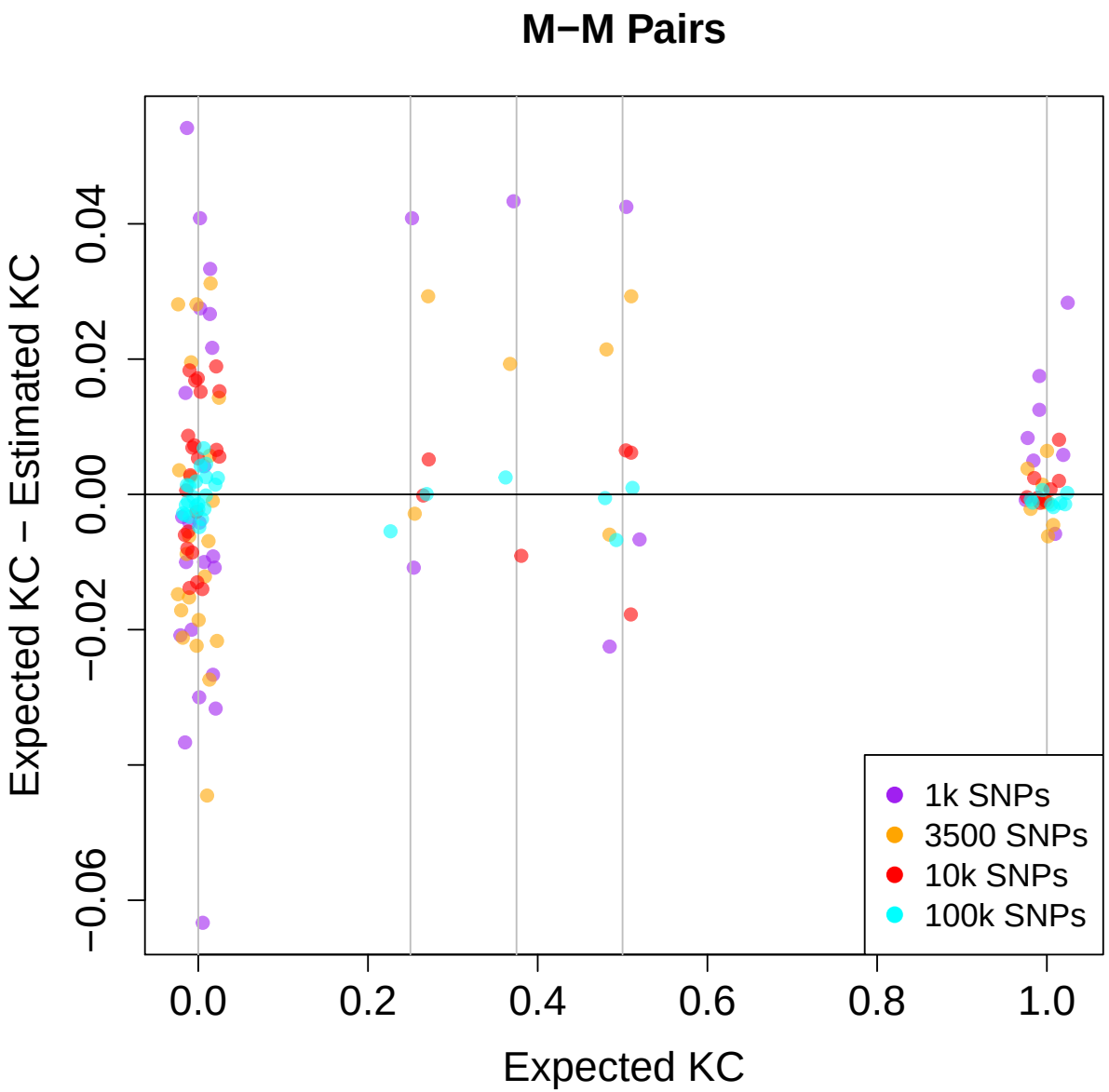

All Pairs
Theoretical KC – Estimated KC
0.04
0.02
0.00
−0.02
−0.06
0.0
0.2
0.4
0.6
0.8
1.0
Theoretical KC
1k SNPs
3500 SNPs
10k SNPs
100k SNPs


All Pairs
Theoretical KC – Estimated KC
0.04
0.02
0.00
−0.02
−0.06
0.0
0.2
0.4
0.6
0.8
1.0
Theoretical KC
1k SNPs
3500 SNPs
10k SNPs
100k SNPs


All Pairs, Autosomes
Expected KC – Estimated KC
0.04
0.02
0.00
−0.02
−0.04
0.0
0.1
0.2
0.3
0.4
0.5
Expected KC
1k SNPs
3500 SNPs
10k SNPs
100k SNPs

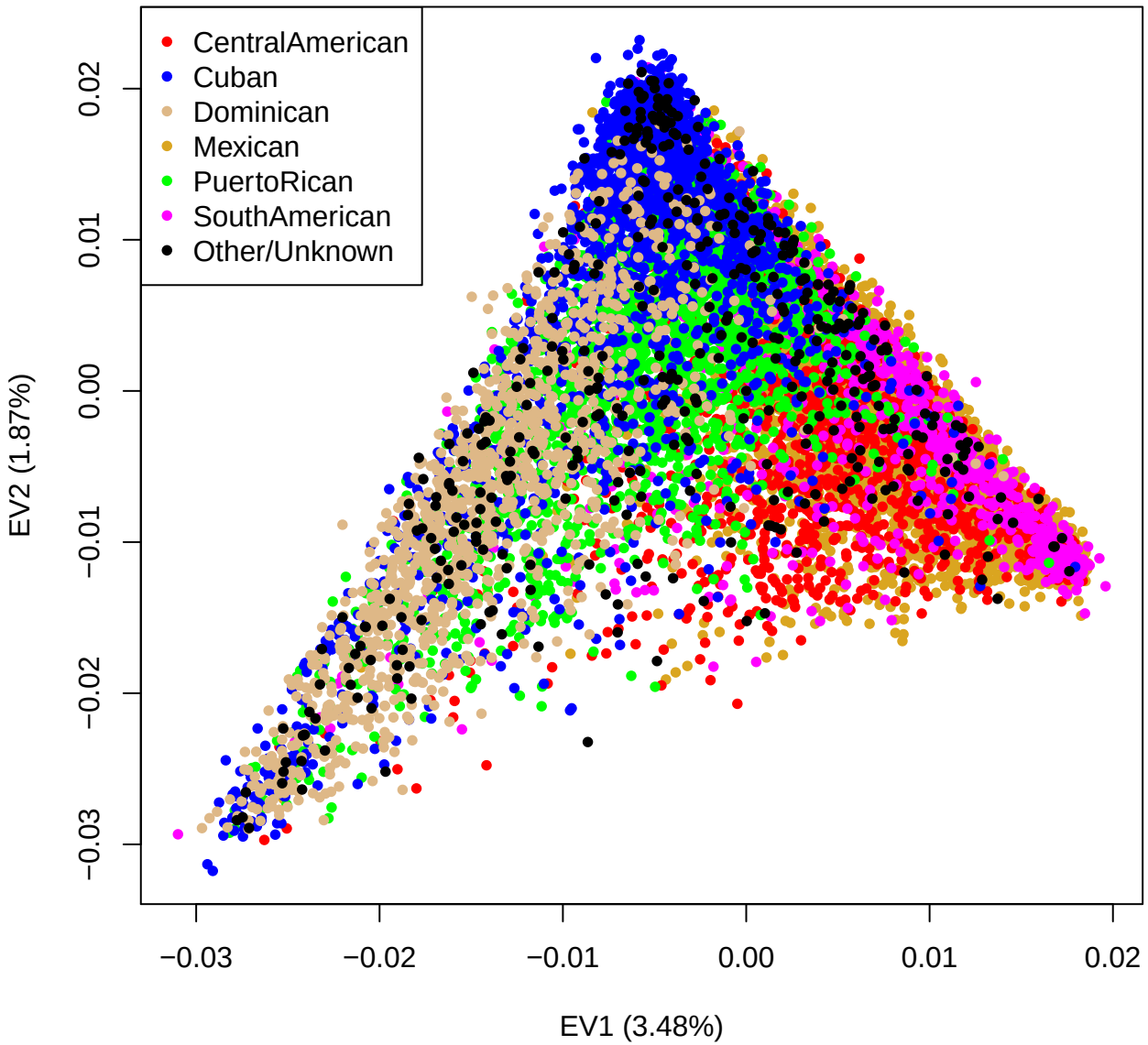

F−F Pairs
1k SNPs
3500 SNPs
10k SNPs
100k SNPs
Expected KC − Estimated KC
0.04
0.02
0.00
−0.02
−0.04
0.0
0.1
0.2
0.3
0.4
0.5
Expected KC

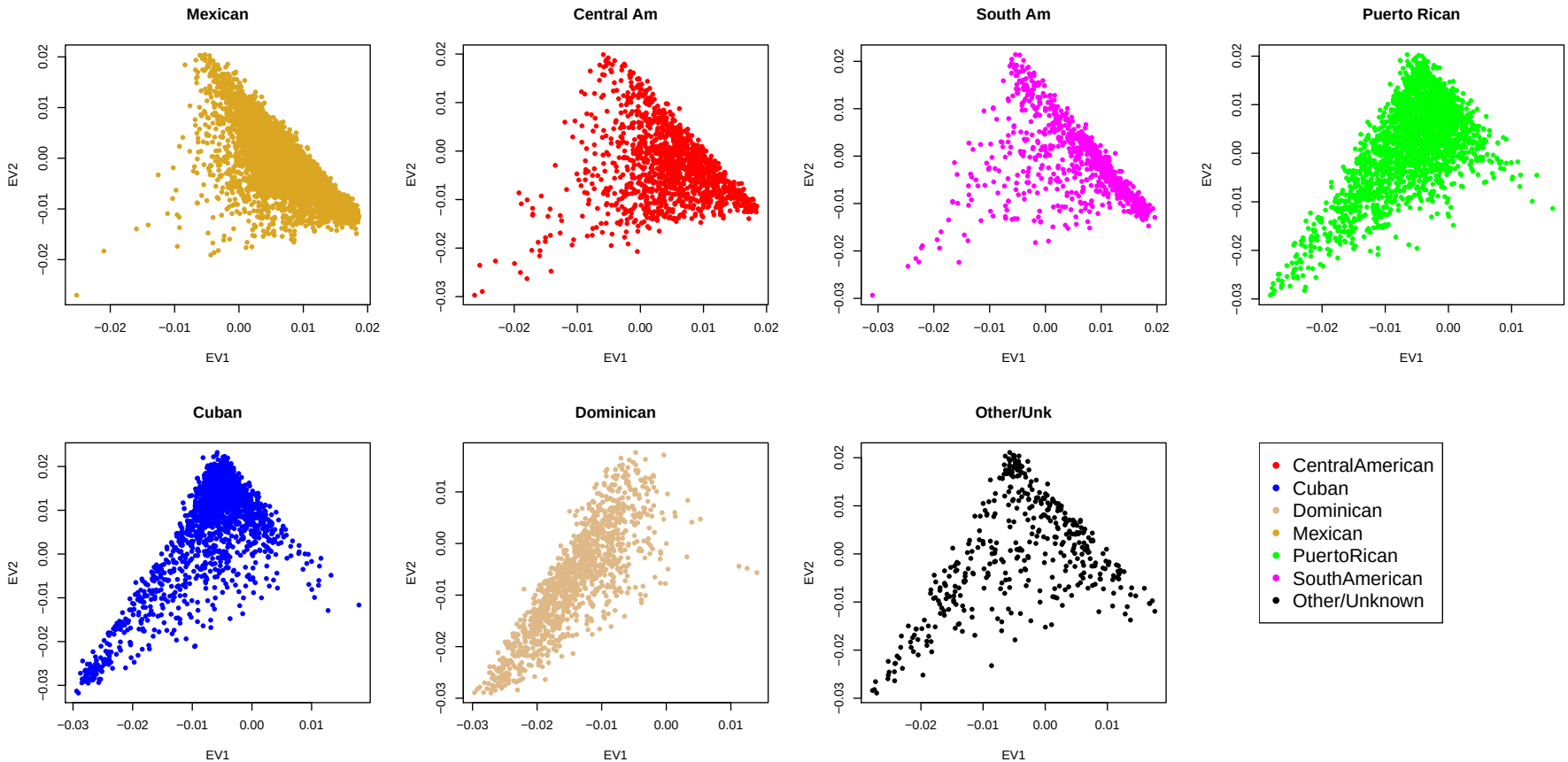

F−M Pairs
Expected KC − Estimated KC
0.04
0.02
0.00
−0.02
−0.04
−0.06
0.0
0.1
0.2
0.3
0.4
0.5
Expected KC
1k SNPs
3500 SNPs
10k SNPs
100k SNPs

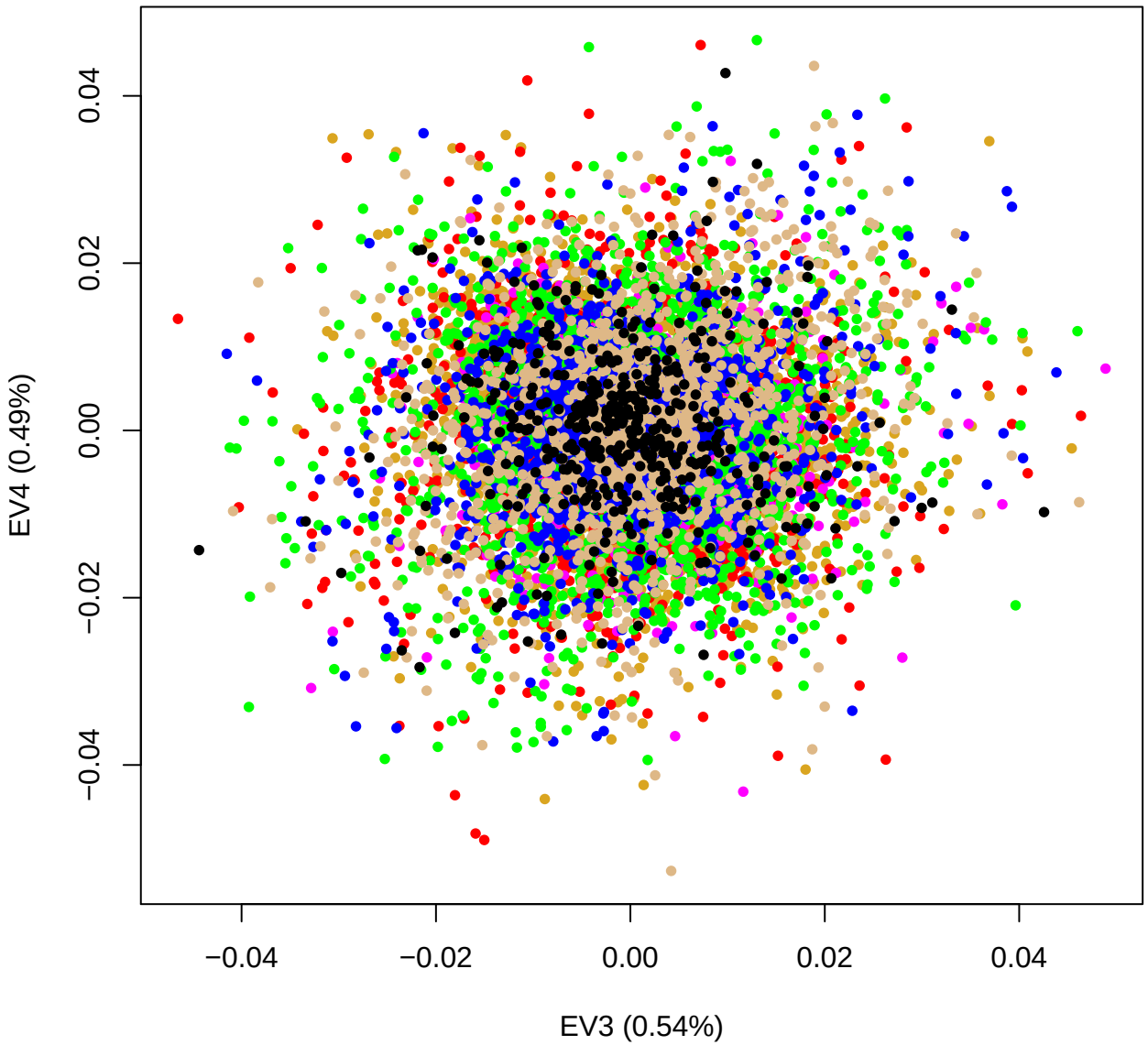

M−M Pairs
Expected KC − Estimated KC
0.04
0.02
0.00
−0.02
−0.06
0.0
0.2
0.4
0.6
0.8
1.0
Expected KC
1k SNPs
3500 SNPs
10k SNPs
100k SNPs

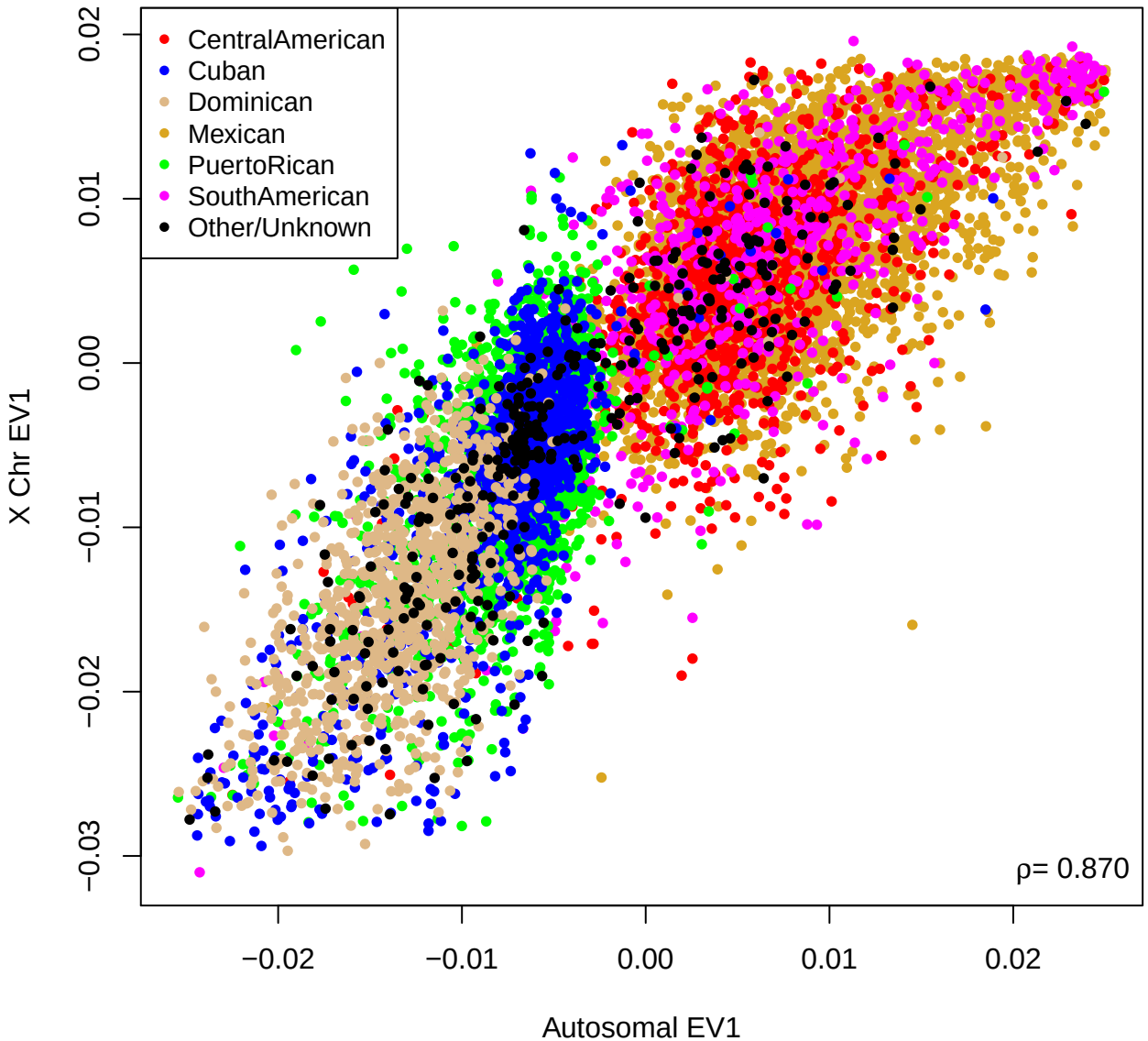
- What about samples from an admixed population?
- We need to determine if we can find 'ancestry adjusted' relatedness estimates on the X chromosome.
  - Using global ancestry, such as PCs estimated on the X.
  - Using local ancestry, such as average X chromosome local ancestry. {we need X chr local ancestry estimates}
  - Or, could we use IBD estimates on the X chromosome from BEAGLE?

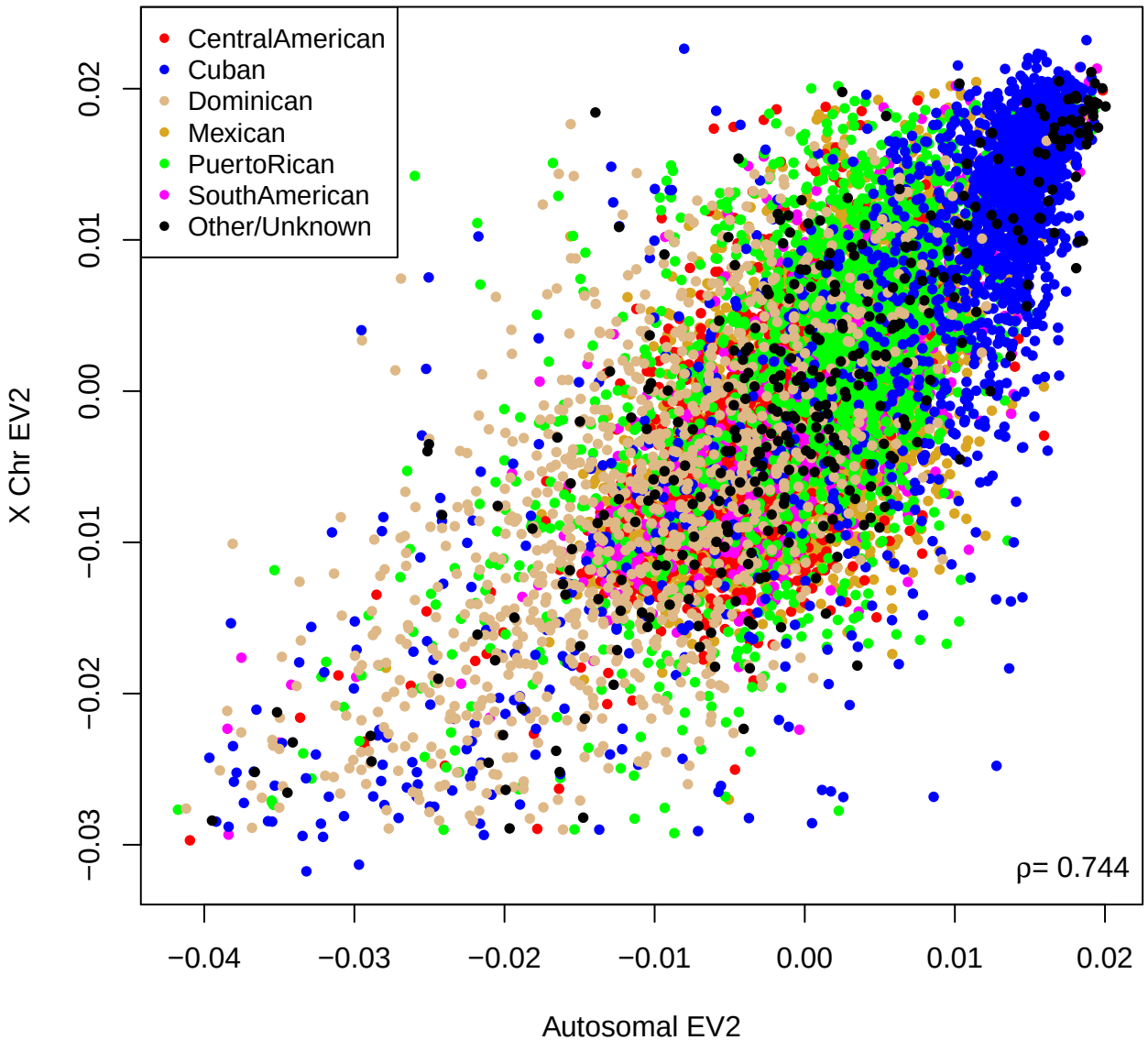
- We estimated PCs in the SOL subjects using 3,600 LD-pruned X chromosome SNPs and PC-AiR.
- The unrelated set `unrelated.pcair.deg4` of 10,272 samples as defined from the autosomes was set, and only study samples (`subj.plink & geno.cntl==0`) excluding `gengrp6.outliers` were projected for a total of 12,747 samples.


CentralAmerican
Cuban
Dominican
Mexican
PuertoRican
SouthAmerican
Other/Unknown
EV2 (1.87%)
0.02
0.01
0.00
−0.01
−0.02
−0.03
−0.03
−0.02
−0.01
0.00
0.01
0.02
EV1 (3.48%)

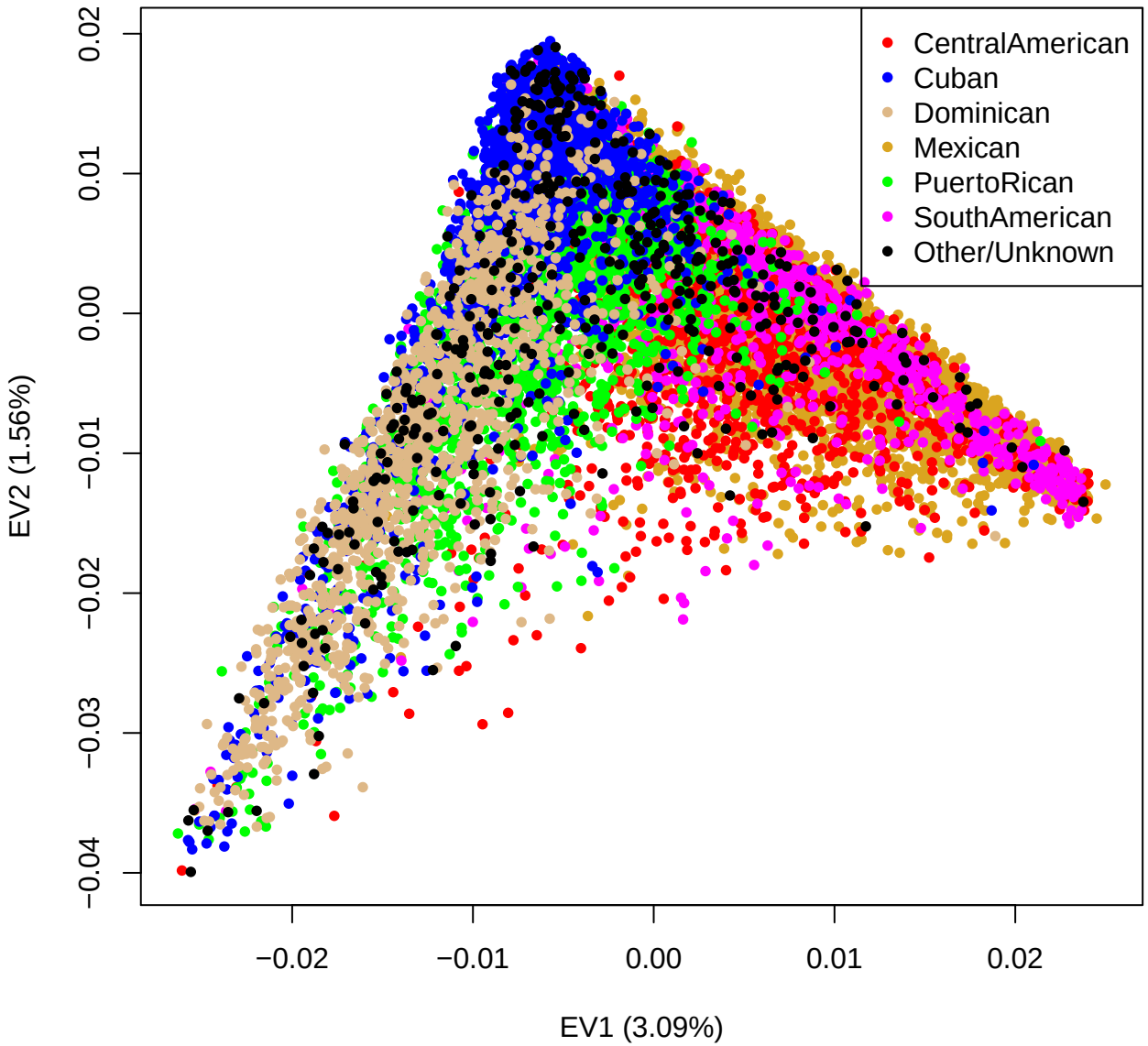

Mexican
Central Am
South Am
Puerto Rican
Cuban
Dominican
Other/Unk
EV1
EV2
CentralAmerican
Cuban
Dominican
Mexican
PuertoRican
SouthAmerican
Other/Unknown

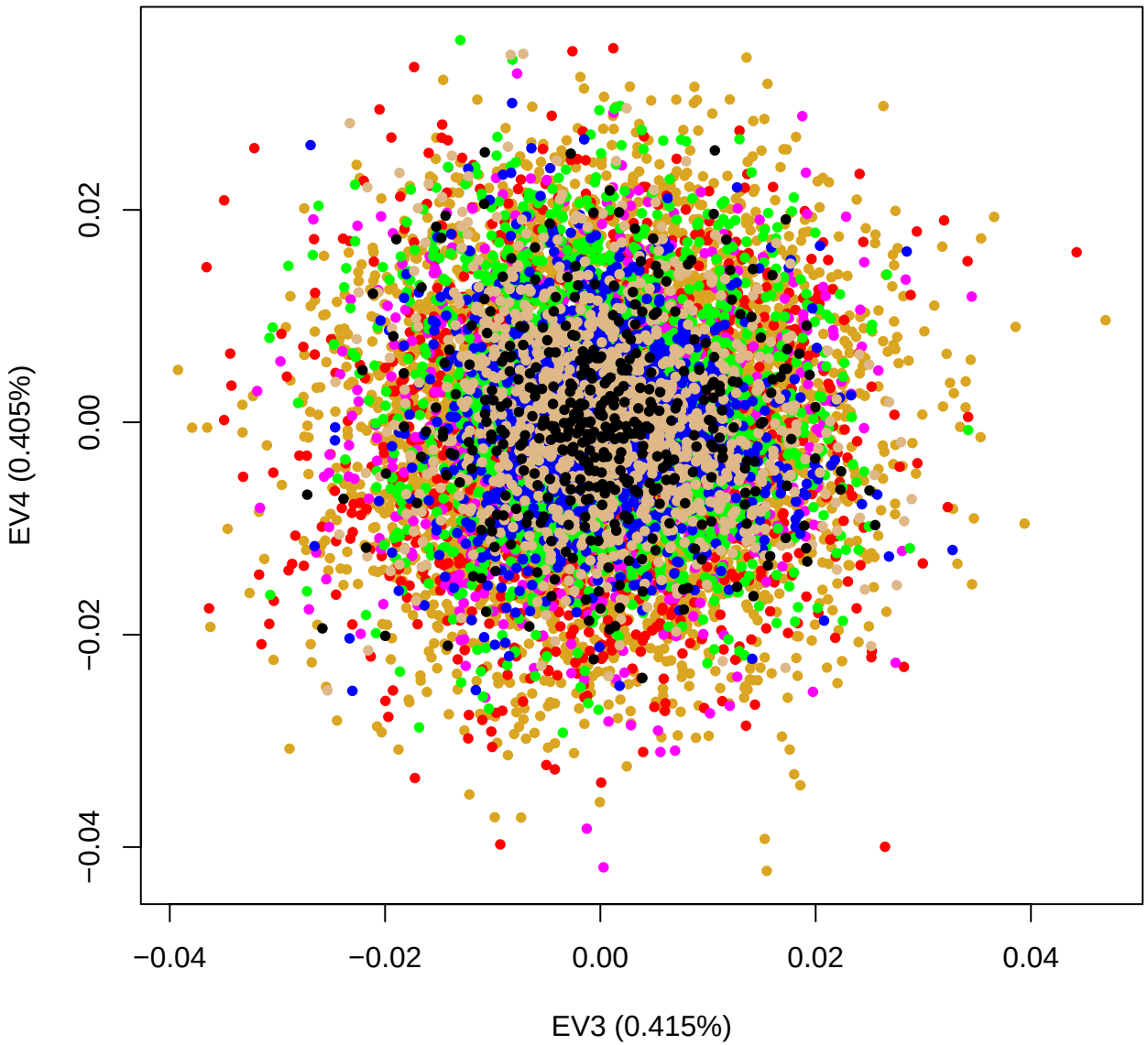

EV4 (0.49%)
0.04
0.02
0.00
−0.02
−0.04
−0.04
−0.02
0.00
0.02
0.04
EV3 (0.54%)

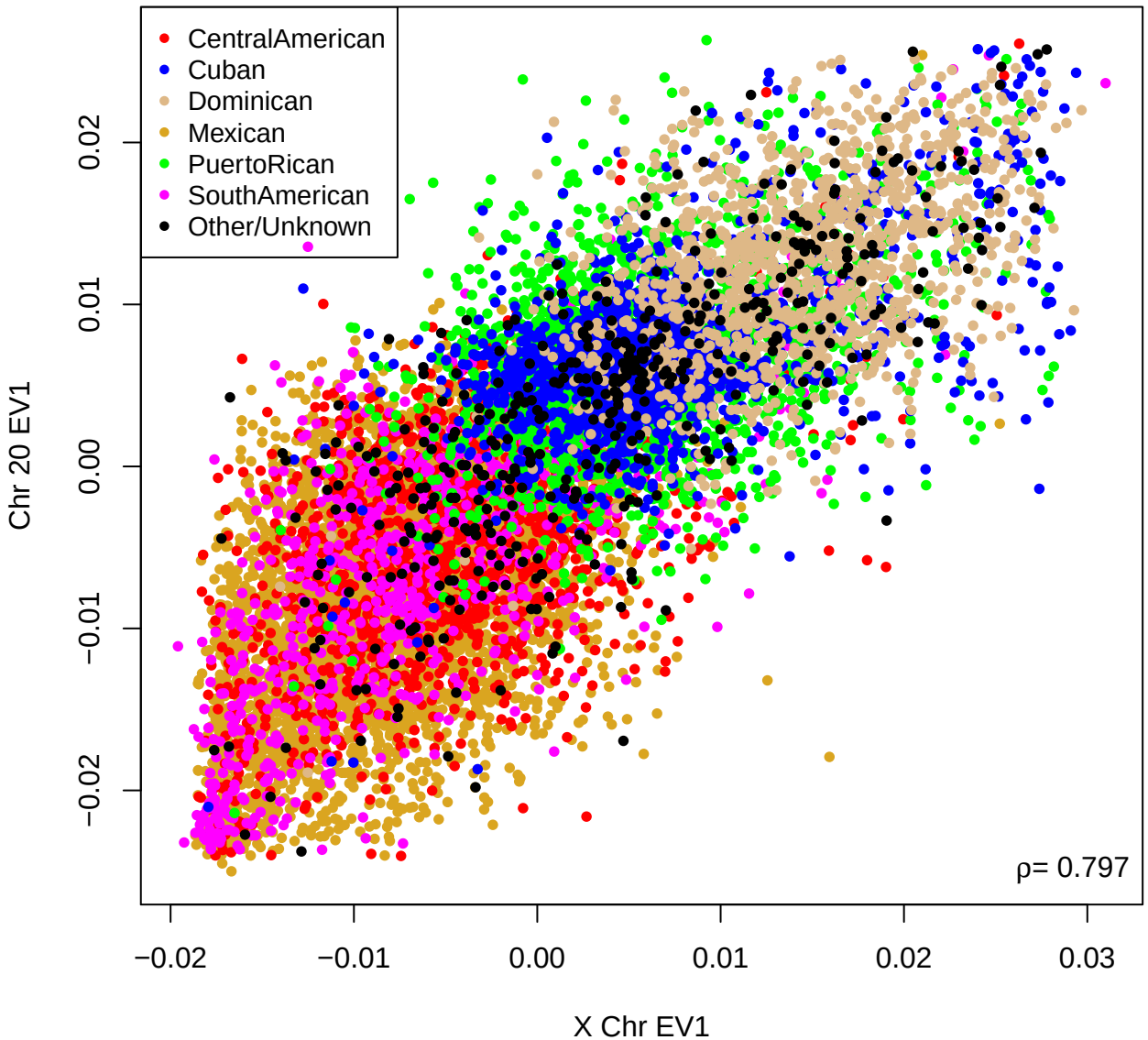

CentralAmerican
Cuban
Dominican
Mexican
PuertoRican
SouthAmerican
Other/Unknown
ρ= 0.870
X Chr EV1
Autosomal EV1
0.02
0.01
0.00
−0.01
−0.02
−0.03
−0.02
−0.01
0.00
0.01
0.02

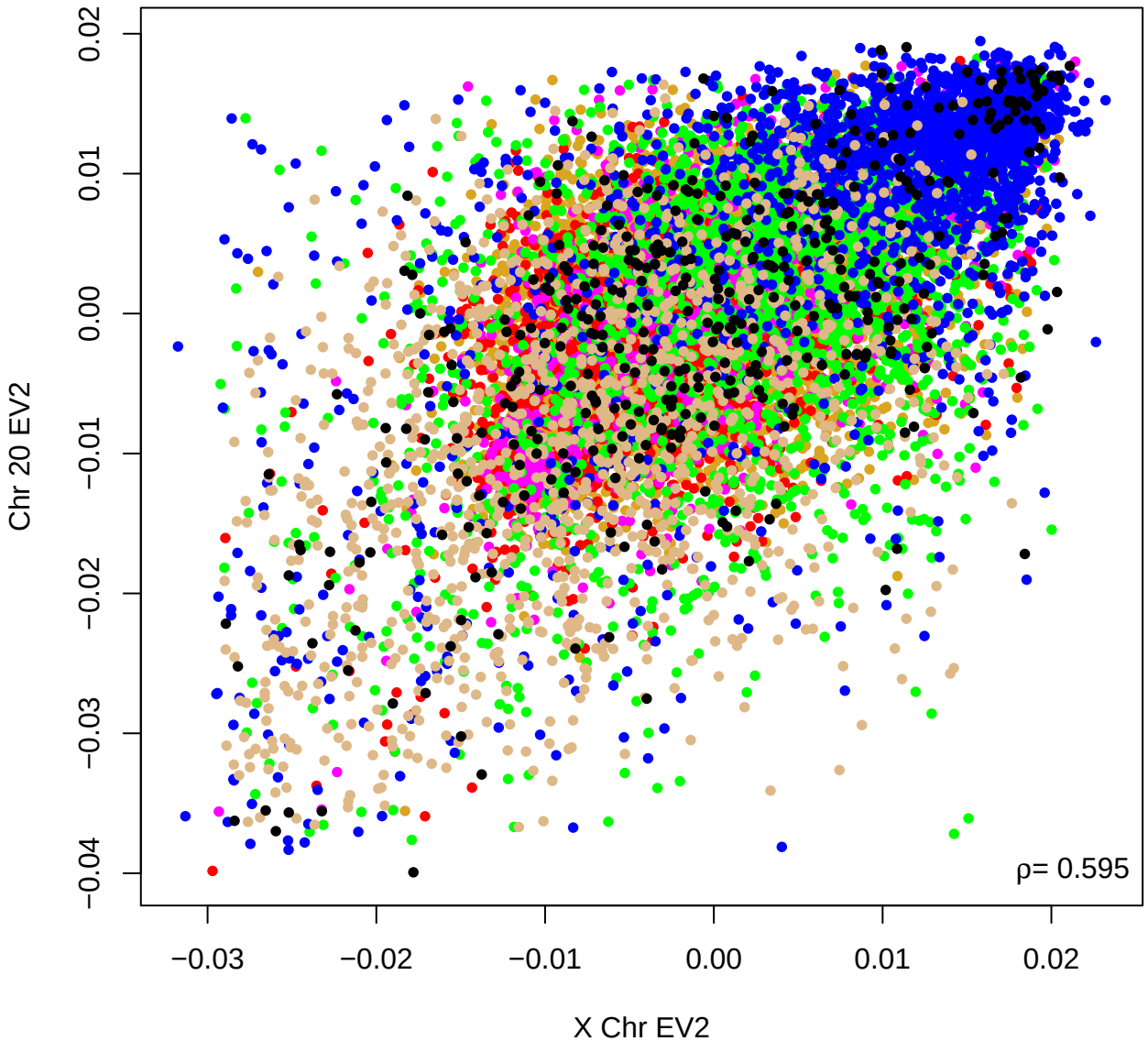

CentralAmerican
Cuban
Dominican
Mexican
PuertoRican
SouthAmerican
Other/Unknown
ρ= 0.744
X Chr EV2
Autosomal EV2
0.02
0.01
0.00
−0.01
−0.02
−0.03
−0.04
−0.03
−0.02
−0.01
0.00
0.01
0.02

- We compare these results which use only 3,600 X chromosome SNPs to a pruned set of 4,413 chromosome 20 SNPs.


CentralAmerican
Cuban
Dominican
Mexican
PuertoRican
SouthAmerican
Other/Unknown
EV2 (1.56%)
EV1 (3.09%)
0.02
0.01
0.00
−0.01
−0.02
−0.03
−0.04
−0.02
−0.01
0.00
0.01
0.02


EV4 (0.405%)
0.02
0.00
−0.02
−0.04
−0.04
−0.02
0.00
0.02
0.04
EV3 (0.415%)


CentralAmerican
Cuban
Dominican
Mexican
PuertoRican
SouthAmerican
Other/Unknown
ρ= 0.797
Chr 20 EV1
X Chr EV1
0.02
0.01
0.00
−0.01
−0.02
−0.02
−0.01
0.00
0.01
0.02
0.03

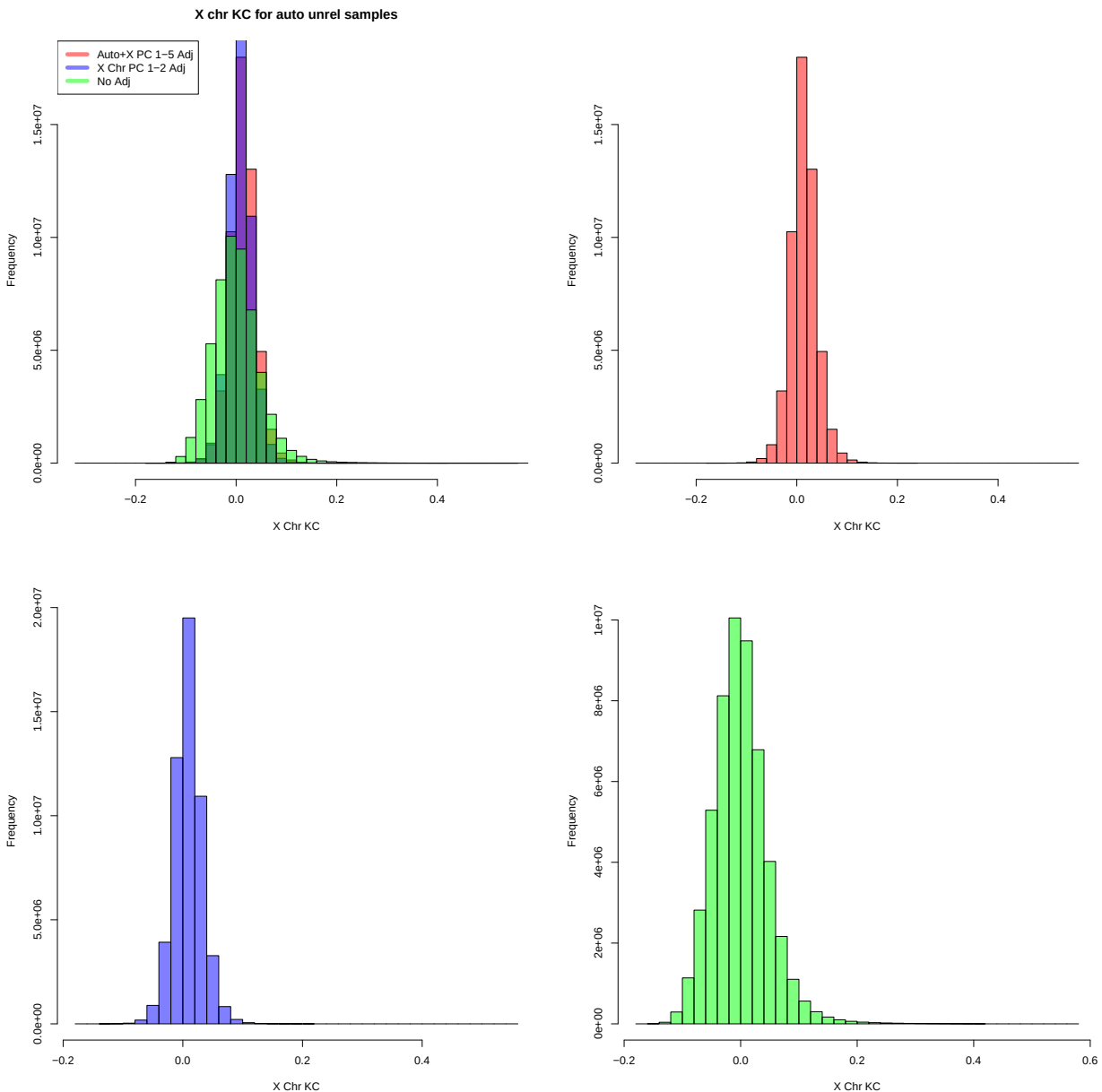

Chr 20 EV2
X Chr EV2
ρ= 0.595
0.02
0.01
0.00
−0.01
−0.02
−0.03
−0.04
−0.03
−0.02
−0.01
0.00
0.01
0.02

We estimated $\mathbf{\Phi_X}$ for all OLGA samples using the following scenarios:

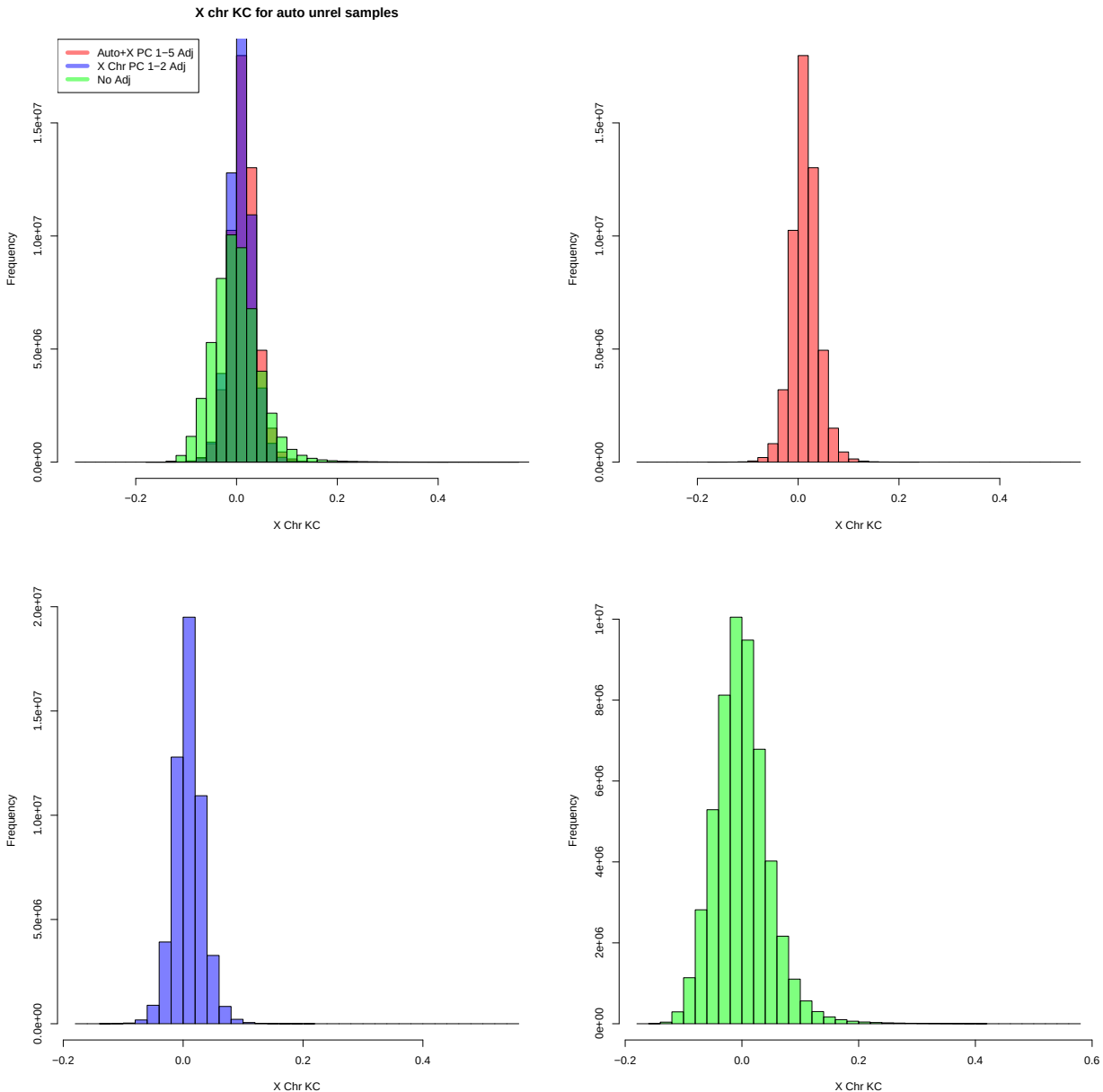
	SNP set	PCA run	EVs used
1	all X chr	-	-
2	all X chr	autosomes + X chr	1-5
3	all X chr	autosomes	1-5
4	all X chr	X chr	1-2
5	pruned X chr	-	-
6	pruned X chr	autosomes + X chr	1-5
7	pruned X chr	autosomes	1-5
8	pruned X chr	X chr	1-2
9	pruned autosomal	autosomes	1-5

All settings used the autosomal unrelated set of 10,272 samples and estimated $\mathbf{\Phi_X}$ for 12,734 study samples posted to dbGaP with no X chromosome anomalies.

Estimate of Φ_X in 10,272 autosomal-unrelated samples models 1-4.



Estimate of Φ_X in 10,272 autosomal-unrelated samples models 5-8.



- What covariates should we include in the model?
- Should we include PCs calculated across the autosomes?
- Should we include PCs calculated on the X chromosome? ADMIXTURE estimates from the X chromosome? Local ancestry estimates averaged across the X chromosome?

- What happens if we simply fit our usual autosomal model when testing X chromosome markers?

- Simulate X chromosome genotypes for 8,000 samples = 500 iterations of the 16-person pedigree; 2,500 unrelated + 5,500 relatives.
- *Note* these samples are from a homogeneous population.
- Simulate phenotypes that follow the model

$$
\begin{aligned}
Y &= \beta_0 + \beta_1 \text{SNP}_X + g_A + g_X + \epsilon \\
g_A &\sim MVN(0, \sigma_A^2 \mathbf{\Phi_A}) \\
g_X &\sim MVN(0, \sigma_X^2 \mathbf{\Phi_X}) \\
\epsilon &\sim MVN(0, \mathbb{I})
\end{aligned}
$$

  where $\sigma_A^2 = 0.3$, $\sigma_X^2 = 0.8$ and for various β_1 values.
- Fit three models:

$$
\begin{aligned}
Y &= \beta_0 + \beta_1 \text{SNP}_X + g_A + g_X + \epsilon \\
Y &= \beta_0 + \beta_1 \text{SNP}_X + g_X + \epsilon \\
Y &= \beta_0 + \beta_1 \text{SNP}_X + g_A + \epsilon
\end{aligned}
$$

α	Adj for X + Auto	Adj for X	Adj for Auto
0.05	0.04983	0.04867	0.07325
0.01	0.00931	0.00913	0.02080
0.005	0.00494	0.00534	0.01095
0.001	0.00094	0.00102	0.00191

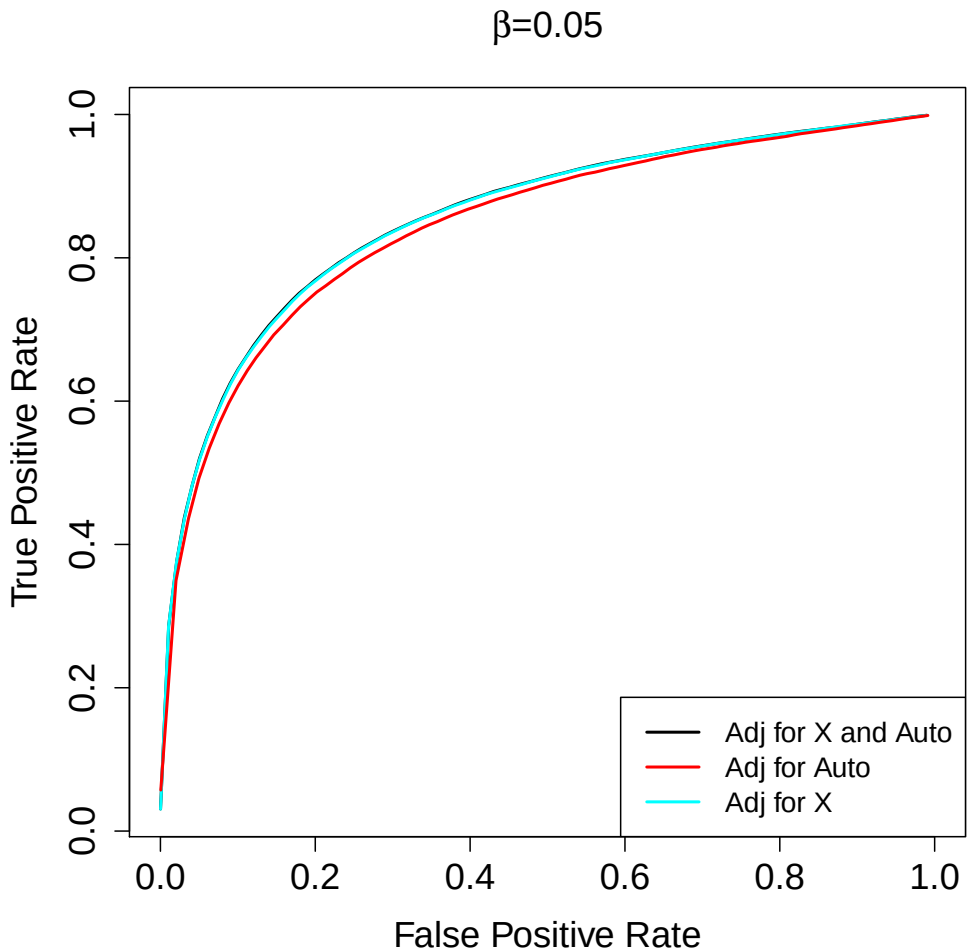
Table: Type I error rate as calculated from 22,000 simulation iterations where β_1=0.

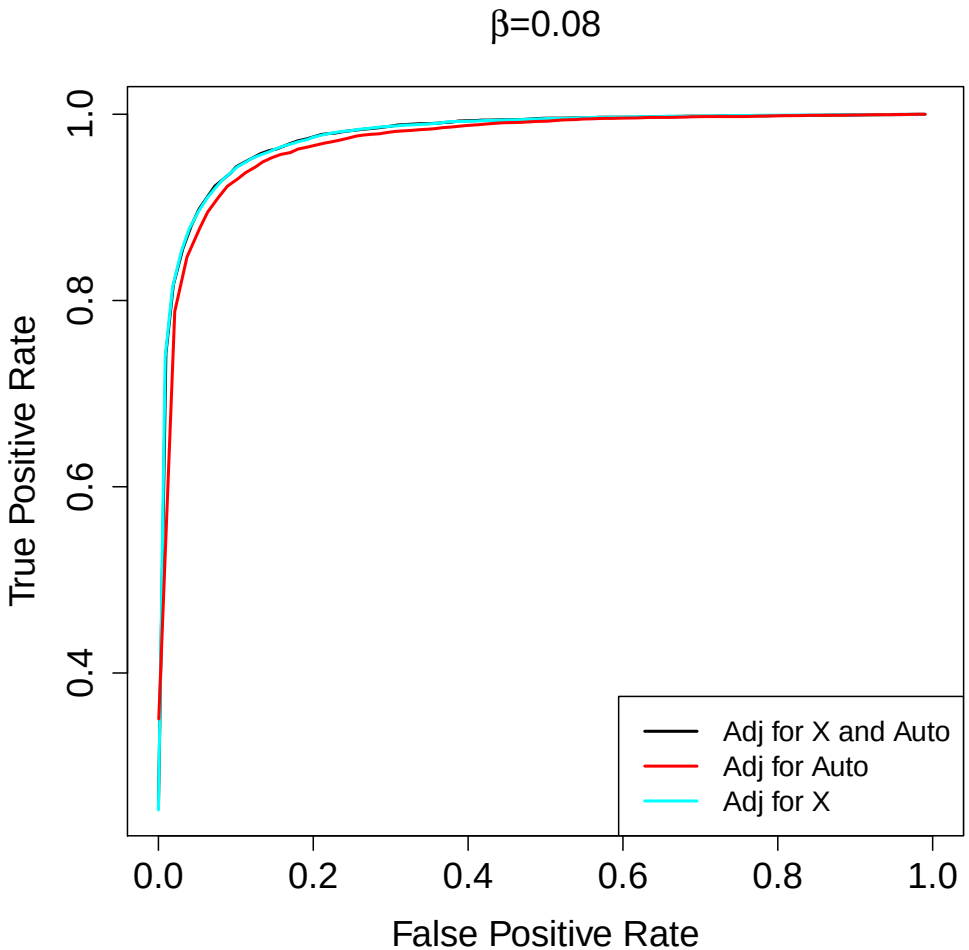
- When only adjusting for autosomal effects, no type I error CIs include the α under consideration.

α	Estimate	95% CI
0.05	0.07325	(0.07040, 0.07610)
0.01	0.02080	(0.01949, 0.02210)
0.005	0.01095	(0.01003, 0.01188)
0.001	0.00191	(0.00150, 0.00233)

Table: Estimate and 95% CI for type I error rate as calculated from 22,000 simulations where β_1=0 and fitting the model $Y = \beta_0 + \beta_1 \text{SNP}_X + g_A + \epsilon$.

- For each α value ranging from 1e-04 to 1, and 7,500 simulation iterations, we calculate the true and false positive rate for each model.




β=0.08
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Adj for X and Auto
Adj for Auto
Adj for X

Supplementary Slides

$$
\begin{aligned}
cov(\mathrm{F},\mathrm{M}) &= \mathbb{E}(\mathrm{F},\mathrm{M}) - \mathbb{E}(\mathrm{F})\mathbb{E}(\mathrm{M}) \\
&= \mathbb{E}(\mathrm{FM}|\mathrm{IBD})\Phi_X + \mathbb{E}(\mathrm{FM}|\text{not IBD})(1-\Phi_X) - (2p)^2 \\
&= [4p^2 + 4p(1-p)]\Phi_X + [4p^3 + 4p^2(1-p)](1-\Phi_X) - 4p^2 \\
&= 4p(1-p)\Phi_X
\end{aligned}
$$

Type I Error CIs

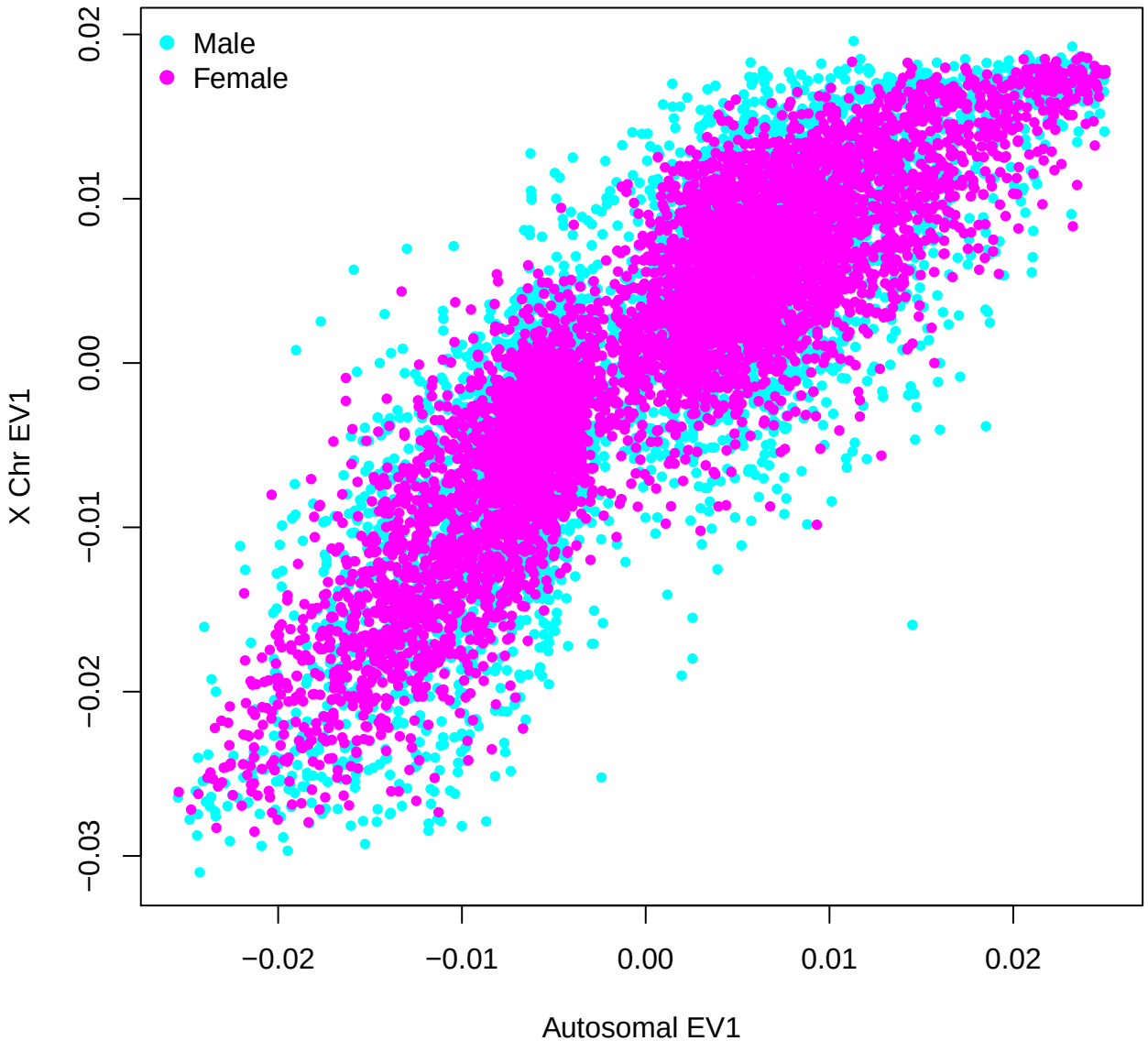
α	Estimate	95% CI
0.05	0.04867	(0.04582, 0.05152)
0.01	0.00913	(0.00783, 0.01043)
0.005	0.00534	(0.00442, 0.00627)
0.001	0.00102	(0.00061, 0.00144)

Table: Estimate and 95% CI for type I error rate as calculated from 22,000 simulations where β_1=0 and fitting the model $Y = \beta_0 + \beta_1 \text{SNP}_X + g_X + \epsilon$.

Type I Error CIs

α	Estimate	95% CI
0.05	0.04983	(0.04698, 0.05268)
0.01	0.00931	(0.00801, 0.01061)
0.005	0.00494	(0.00402, 0.00587)
0.001	0.00094	(0.00052, 0.00135)

Table: Estimate and 95% CI for type I error rate as calculated from 22,000 simulations where β_1=0 and fitting the model $Y = \beta_0 + \beta_1 \text{SNP}_X + g_A + g_X + \epsilon$.


Male
Female
X Chr EV1
Autosomal EV1
0.02
0.01
0.00
−0.01
−0.02
−0.03
−0.02
−0.01
0.00
0.01
0.02