Brawn and Brains: a Robust and Powerful approach to X-inclusive Whole-genome Association Studies

Bo Chen*, Radu V. Craiu, and Lei Sun*****

Department of Statistical Science, University of Toronto, ON, M5S 3G3 , Canada

**email:* bochen@utstat.toronto.edu

***email:* craiu@utstat.toronto.edu

****email:* sun@utstat.toronto.edu

SUMMARY: X-chromosome is often excluded from whole-genome association studies due to a number of complexities. Some are apparent, e.g. sex-specific allele frequencies, sex-gene interaction effects, and the choice of (additive or other) genetic models, while others are subtler, e.g. random, skewed or no X-inactivation, and the choice of risk allele. In this work, we aim to consider all these complexities jointly and propose a regression-based association test. We provide theoretical justifications for its robustness in the presence various aforementioned model uncertainties, as well as for its improved power under certain alternatives as compared with existing approaches. For completeness, we also revisit the autosomes and show that the proposed framework leads to a robust and sometimes much more powerful test than the standard method. Finally, we provide supporting evidence from simulation and application studies.

KEY WORDS: Model selection; Confounding; Genome-wide association studies; X-chromosome.

1. Introduction

In genome-wide association studies (GWAS) and next generation sequencing (NGS) studies, X-chromosome has been often excluded due to its complexity compared to autosomes. Wise et al. (2013) found that for every GWAS paper published from January 2010 to December 2011 and included in the NHGRI GWAS catalog, "*only 33% (242 out of 743 papers) reported including the X-chromosome in analyses*". There are many analytical challenges related to X-inclusive association studies. Some are for both autosomes and X-chromosomes, and some are specific to X-chromosomes.

Throughout this paper, we use Y to denote phenotype or outcome of interest, which could be binary or continuous, and G to denote genotype of a single nucleotide polymorphism (SNP). A single SNP has two alleles: r and R, one of which is the risk allele with allele frequency p and the other is reference allele. It needs to be noted that the major allele could be risk allele and p is not necessarily the minor allele frequency (MAF) less than 0.5. An autosome SNP has three genotypes, namely rr, rR and RR. Coding of G for each genotype could be $G_A = (0, 1, 2)$ for additive effect, and $G_D = (0, 1, 0)$ for dominant effect. An X-chromosome SNP has five genotypes, rr, rR and RR for females and r and R for males. We will discuss the coding of G in more details below. The main question of phenotype-genotype association analysis is to test $H_0 : Y$ is not associated with G. In addition, we use S to represent sex-specific effect and Es to represent other environmental effects. When both effects exist, there may also exist $G \times S$: genotype-sex interaction, $G \times E$: gene-environmental interaction and $S \times E$: sex-environmental interaction. β (with corresponding subscripts) denotes the effect sizes of each covariate.

For any statistical approaches focusing on X-chromosome analysis, we summarize 8 major challenges that must be properly addressed. As we discuss below, challenges C1 to C3 are genome-wide, and C4 to C8 are specific to X-chromosome.

- C1: Quantitative vs. binary traits/phenotypes
- C2: Genotype based vs. allele based
- C3: Additive vs. genotypic model (with dominant term)
- C4: Sex S as a covariate must be included or not
- C5: Genotype-sex interaction $G \times S$ should be included or not
- C6: X-chromosome inactivation (XCI) vs no inactivation
- C7: If XCI, the inactivation is random vs. skewed
- C8: Reference allele R vs. r

C1 and C2. Classic allele based tests from case-control studies require binary phenotype data so that the Pearson chi-squared test statistics can be computed by contingency tables. HardyWeinberg equilibrium (HWE) assumption must also be met to achieve correct type I errors. For quantitative phenotypes and any departure from HWE, the most commonly used approaches are genotype based tests under regression models. Regression models support various types of phenotype data and HWE assumption is not required. Sasieni (1997) had a detailed discussion about allele based tests with HWE assumption versus genotype based tests. Another reason in favor of regression model is that additional covariates such as environmental factors can be easily incorporated in the model.

C3. The genotype-based tests require a correct assumption of the genetic model, which has been a long standing controversy. For both autosome and X-chromosome SNPs, the genetic model either assumes a specific relationship between the effects of rr, rR and RR, such as recessive, additive, dominant, multiplicative, or assumes no specific relationship between each genotype (genotypic model), where the total genetic effects are decomposed as a combination of additive and dominant effect. Each assumption leads to a different model, and Bagos (2013) had a good review paper of several model selection approaches. When the true genetic model is unknown, the main idea is to combine each test statistic or p-value

under different models. However, the way to combine these tests are quite ad hoc, and it is lack of theoretical justifications that how and why they should be combined.

On the other hand, a common practice for simplicity is to only examine additive models, as the additive model has reasonable power to detect both the additive and dominant effects (Bush and Moore, 2012). In addition, Hill et al. (2008) have shown that additive variance typically accounts for over half and often close to 100% of total genetic variance, even if there are non-additive effects at the level of gene action. It needs to be noted that people are usually reluctant to assume the genotypic model. Although it is the most general assumption, the test is believed to be less powerful due to the extra degree of freedoms of the test statistics. However, we find such belief is not necessary correct in the context of GWAS. We derive the upper bound of the power loss by incorporating the other covariate for dominant effect and compare to the potential power gain, and find it may be worth to allow both the additive and dominant effects in the model.

C4 and C5. The other challenges are specific to X-chromosome, due to the fundamental differences between females and males. First, sex-specific effects may exist in biological point of view. Next, sex is a classic confounder associated to both the genotype and the phenotype. If covariate S is not included in the model, the type I error for testing the genotype effect can be inflated. Ozbek et al. (2018) has extensitve simulation studies to show the type I error inflation. Furthermore, different effect sizes of the same SNP in females and males are recognized as genotype-sex interaction effects. Proper tests allowing for interaction effects need to be developed.

C6 and C7. The next complications relate to the uncertainty of the biological status of X-chromosome SNPs. X-chromosome inactivation is the phenomenon that one of the two alleles in females is selected to be silenced, so that the effects of female genotypes may be reduced. In brief, the additive coding of rr, rR and RR becomes 0, 0.5 and 1 rather than 0, 1 and 2.

The challenge is that although we know about 15% of genes on X-chromosome are escaped from XCI at population level (Carrel and Willard, 2005), we are uncertain if XCI occurs or escapes on each SNP. Even though we are certain that XCI occurs on one particular SNP, at individual level it is still unknown which allele is inactivated. Wang et al. (2014) discussed various studies suggesting a biological plausibility of skewed inactivation so that one allele is more likely to be inactivated than the other, while the additive model in essence assumes two alleles have equal probability of inactivation.

C8. Lastly, when allele frequency difference is significant, females and males may have different minor alleles. For autosome SNPs, people usually choose the minor allele with allele frequency less than 0.5 as the risk allele, because switching the risk allele and reference allele does not change the statistical inference and thus choosing an arbitrary risk allele does not cause a problem. However, for X-chromosome SNPs, switching the reference allele and risk allele may lead to different statistical models and yield different inferences. When minor allele is different for females and males, the risk allele may be unknown and it becomes a challenge to choose the risk allele. It needs to be noted that sex-stratified tests may not solve the challenge, because stratification by sex may result in considerable loss of power (Clayton, 2008), especially when allele frequency difference is significant for females and males.

We summarize the genotype codings after considering all X-specific challenges in Table 1. If the risk allele and XCI status are both unknown, there are $2 \times 2 = 4$ ways to code the additive covariates G_A, and 2 ways to code the genotype-sex interaction GS. We will discuss in section 3 that skewed inactivation can be represented by the dominant effect coding G_D.

[Table 1 about here.]

In recent years, quite a few methods have been proposed for X-chromosome association studies. Zheng et al. (2007) proposed a few tests without considering X-chromosome inactivation. In contrast, Clayton (2008, 2009) discussed analytical strategies assuming X-

chromosome is always inactivated. Hickey and Bahlo (2011) and Loley et al. (2011) separately performed simulation studies and gave a thorough comparison of Zheng et al. and Clayton's tests. Based on these simulation studies, Konig et al. (2014) provided a detailed guideline for including X-chromosome in GWAS. The problem is they suggested different tests under different assumptions of genetic model, interaction effects, XCI status and so on, and it is not always possible to check these assumptions in practice. Gao et al. (2015) developed a software toolset for X-chromosome association studies. Recently, Zhongxue et al. (2017) improved existing sex-stratified tests by eliminating assumptions of genetic models, but they still needed to assume same risk allele for females and males, and sex-genotype interaction effects could not be measured. Focusing on XCI status, Wang et al. (2014) proposed a maximum likelihood solution to handle the uncertainty of XCI as well as skewed inactivation, and provided an XCI model selection method in their most recent paper (Wang et al., 2017). In addition, Chen et al. (2017) used Bayesian model averaging (BMA) method to solve XCI uncertainty. However, both approaches only considered the additive model, and it is unclear how to include non-additive covariates in regression analysis with unknown XCI status. Furthermore, both approaches were only illustrated by simulation studies, and it would be more appealing to derive a theoretical justification.

After reviewing all up-to-date methodology developments on X-chromosome association studies, we believe there is currently no approach which can handle all the 8 challenges discussed above simultaneously. The target of this paper is to propose a theoretically justified robust method that can solve all these challenges in most general framework, while the test powers are well maintained and even improved in most practical situations. The proposed tests are based on regression models, which allow for both quantitative and binary phenotypes as the response variables, departure from HWE and incorporating extra covariates. In section 2, we discuss the long-lasting controversy between additive models and genotypic models.

We revisit autosome SNPs for better illustrating of the benefits of genotypic model, which leads to a robust and sometimes much more powerful test than additive model. In section 3, we propose our main theory to address the challenges specific to X-chromosome. Section **??** provides supporting evidence to our proposed approach from application studies. Finally, we discuss the limitations of our approach and possible future work in section 4.

2. Additive vs. genotypic models

2.1 *Theory of chi-squared distributions*

For completeness and a more clear demonstration of the model selection challenge, we first revisit autosome studies. We find that in general, Hill et al. (2008)'s result does not warrant the exclusion of dominant covariate G_D in regression model. Although the additive effect may account for the majority of total genetic effect, excluding G_D does not necessarily increase test power. In order to determine whether G_D needs to be included or not, two questions must be answered. First, when all genetic effect is additive, what is the power loss by introducing the extra covariate G_D? Second, when there exists some dominant effect, will the power increase or decrease by introducing G_D?

To answer these two questions, we first define the additive model and genotypic model in generalized linear regression framework. Let g be the link function. The additive model is defined by

$$E[g(Y)] = \beta_0 + \beta_A G_A,$$

and the genotypic model is defined by

$$E[g(Y)] = \beta_0 + \beta_A G_A + \beta_D G_D,$$

Although the HWE assumption is not required, we adopt it only for the purpose of simplifying the computation. The three genotype groups, rr, rR and RR have frequencies $(1-p)^2$, $2p(1-p)$ and p^2, and G_A and G_D are coded as 0, 1, 2 and 0, 1, 0 correspondingly. Covariates

for environmental factors can be freely added to both models. For notation simplicity we denote the additive model by $Y \sim G_A$ and genotypic model by $Y \sim G_A + G_D$ below.

We want to compare two tests: $H_0 : \beta_A = 0$ under the additive model and $\beta_A = \beta_D = 0$ under the genotypic model. The test statistics used most often in regression models are Wald, Score and likelihood ratio statistics, which all follow asymptotic non-central chi-squared distribution under the alternative hypothesis. We define two test statistics by $W_1 \sim \chi^2_{(1,ncp_1)}$ and $W_2 \sim \chi^2_{(2,ncp_2)}$, where 1 and 2 denote degree of freedoms, and ncp_1 and ncp_2 are corresponding non-centrality parameters. If the true genotype effect is all additive, then $ncp_1 = ncp_2$. The power difference of W_1 and W_2 depends on both the non-centrality parameters and type I error α. When $ncp_1 = ncp_2 = 0$ or $\alpha = 0$, both tests have no power; when non-centrality parameters are sufficiently large or α close to 1, both powers are close to 1. To achieve the maximum power loss of W_2, we expect a moderate value of both the non-centrality parameter and α. We show the maximum power loss numerically in Web Appendix A, where the maximum power loss is 0.114 when $\alpha = 0.0025$ and $ncp = 10.6$. It implies the power loss of using the genotypic model is capped by 0.114, regardless of type I error level, sample size and size of additive effects. It needs to be noted that although we assume additive model is correct, the maximum power loss is same for all 1 vs 2 degree of freedom models. For instance, if the dominant model is correct, the power loss is still capped by 0.114 by using the genotypic model.

With capped power loss, we want to investigate the power gain by testing W_2 when the true genotype effect is not additive. When dominant effect exists, the non-centrality parameters can be written as $ncp_2 = ncp_1 + \Delta_{12}$, where $\Delta_{12} > 0$. For fixed value of ncp_1, when Δ_{12} is close to 0, we still expect W_2 to be less powerful than W_1. As Δ_{12} increases, there is a threshold value of Δ_{12} which makes W_1 as powerful as W_2. When Δ_{12} is greater than the threshold, W_2 is more powerful, and the power goes up to 1 for large Δ_{12}. Compared to the

maximum power loss, the maximum power gain can be technically as large as $1-\alpha$ when $ncp_1 = 0$ and $\Delta_{12} \to \infty$. In Web Appendix A, we choose a few practical values of ncp_1 and Δ_{12} and plot the test powers of W_1 and W_2. To clearly illustrate the power gain, we assume the worst case (maximum power loss) scenario where $\alpha = 0.0025$. We show when Δ_{12} is as large as ncp_1, the power gain can be much higher than power loss. Therefore, the genotypic model should not be overlooked in association studies with autosome SNPs.

2.2 *Non-centrality parameters and corresponding test power computation*

The above power computation is based on the theoretical values of non-centrality parameters, which must be computed from sample size and genotype effect size under the additive or genotypic model. When the sample size $n \to \infty$, we want each test has a limiting chi-squared distribution, but the non-centrality parameter under alternative hypothesis would move toward infinity for fixed value of $\beta = (\beta_0, \beta_A, \beta_D)$. As in convention, we assume $\beta = c/\sqrt{n}$. Instead of specifying β, we fix the value of constant vector c, so that $\beta \to 0$ and the non-centrality parameter under alternative hypothesis converges to finite number as $n \to \infty$. We provide more discussions about the convergence of asymptotic non-centrality parameters in section 4. We then use standard technique in Cox and Hinkley (1974) to compute the asymptotic non-centrality parameters for the test under genotypic model as described below.

We write the generalized linear models in matrix form: $E[g(Y)] = X\beta$ where X is the design matrix. Suppose we want to test $H_0 : \beta_2 = 0$, where β_2 is a subset of β. To compute the non-centrality parameter, we partition $X = (X_1, X_2)$, $\beta = (\beta_1, \beta_2)$ according to the null hypothesis, and the expected Fisher information matrix of β is partitioned accordingly: $H(\beta_1, \beta_2) = \begin{bmatrix} H_{11}(\beta_1, \beta_2) & H_{12}(\beta_1, \beta_2) \\ H_{21}(\beta_1, \beta_2) & H_{22}(\beta_1, \beta_2) \end{bmatrix}$. Then the non-centrality parameter equals to

$$ncp = \beta_2'[H_{22}(\beta_1, 0) - H_{21}(\beta_1, 0)H_{11}^{-1}(\beta_1, 0)H_{12}(\beta_1, 0)]\beta_2. \quad (1)$$

For genotypic model, $\beta_1 = \beta_0$, $X_1 = 1_n$, and $\beta_2 = (\beta_A, \beta_D)$, $X_2 = (G_A, G_D)$. Specifically,

we derive H under linear model and logistic model and compute corresponding *ncp*. The mathematical details are given in Web Appendix C. Let σ^2 be the variance of the error term in linear model. We show when $\sigma^2 = 4$, linear and logistic model has equal asymptotic non-centrality parameters for same X and β.

The computation of non-centrality parameter under the additive model is less straightforward, because the additive model is indeed misspecified when the dominant effect $\beta_D \neq 0$. Although the derivation is difficult under the canonical parametrization of the genotype as defined above, the result from Begg and Lagakos (1992) implies that a re-parametrization of genotype coding may considerably simplify this derivation. Detailed steps are provided in Web Appendix D.

Once the non-centrality parameters are computed, we may compare test powers of the additive and genotypic model when $\beta_D \neq 0$. Choosing $\alpha = 0.0025$, we consider a realistic situation where $n = 1000, \beta_0 = -0.3, \beta_A = 0.3$ and β_D change from -0.6 to 0.6. We then plot the power of both tests as a function of β_D in Figure 1, which represents both the logistic model and linear model with $\sigma^2 = 4$. Risk allele frequencies are chosen to be 0.2, 0.5 and 0.8.

[Figure 1 about here.]

Figure 1 indicates that the power gain by using the genotypic model can be as much as 0.4 in realistic situations, which is quite significant compared to the maximum power loss of 0.113. In practice, the strength of dominant effect is usually unknown. In such case including the dominant covariate is more like a risk-free solution: without sacrificing much test power, the potential power gain may be significant.

3. X-chromosome Challenges

3.1 *Type I error control and choice of risk allele*

We now consider association analysis on X-chromosome SNPs where the covariates are defined in Table 1. When testing for the genotype effects, we note that they are usually correlated with sex effects. The correlation has two implications. First, when genotype effect exists, sex becomes a confounding variable. The sex effect is hard to explain separately and it in fact helps explaining the genotype effect. Second and more importantly, when sex effect exists but genotype effect does not exist, the correlation will lead to an inflated type I error for testing the genotype effect if sex is not included in the model. Including sex as the covariate warrants the correct type I error for testing the genotype effect. **?** provided extensive simulation studies to show both the type I error inflation and correct type I error control when sex is included, and we would agree with their conclusion that sex should always been included in regression models.

As shown in Table 1, the coding of G_A depends on the risk allele, and model $Y \sim G_A$ may yield different test statistics under different risk allele assumptions. At first sight, it may seem unclear that how we choose the correct way to code G_A when risk allele is unknown. However, with sex as the covariate, there turns out to be a connection between different risk alleles. For instance, we observe two models with no XCI and different risk alleles,

$$E[g(Y)] = \beta_{10} + \beta_{1S}S + \beta_{1A}G_{A,R,N} \text{ and } E[g(Y)] = \beta_{20} + \beta_{2S}S + \beta_{2A}G_{A,r,N}$$

where we want to test $\beta_{1A} = 0$ or $\beta_{2A} = 0$ under each model. We note that $G_{A,r,N} = 2 - G_{A,R,N} - S$, which yields $\beta_{1A} = -\beta_{2A}$, so it is equivalent to test $\beta_{1A} = 0$ and $\beta_{2A} = 0$ under two models, and we can further show that test statistics under two models are exactly equal. It provides us some intuition that the problem of unknown risk allele is solved when sex is included as the covariate because two tests then become indistinguishable. To make our intuition more rigorous, we propose the following theorem:

THEOREM 1: *For vector Y of length n, Let M_1 and M_2 be two generalized linear models with same link function g, $g[E(Y)] = X_1\beta_1$ and $g[E(Y)] = X_2\beta_2$, where X_1, X_2 are $n \times p$ design matrices and β_1, β_2 are vectors of length p. Let $\beta_1' = (\beta_{11}', \beta_{12}')$ and $\beta_2' = (\beta_{21}', \beta_{22}')$, where β_{11} and β_{21} have length $(p-q)$ and β_{12} and β_{22} have length q. If there exists a transformation matrix $T = \begin{pmatrix} T_1 & T_{12} \\ 0 & T_2 \end{pmatrix}$ such that $X_2 = X_1T$, where T_1, T_2 are $(q-p)\times(q-p)$ and $q \times q$ invertible matrix, then the test statistics (Wald, Score or LRT) for testing $\beta_{12} = 0$ and $\beta_{22} = 0$ are equal under the technical assumptions given in Web Appendix B.*

We prove Theorem 1 in Web Appendix B. To make the two test statistics equal, an intuitive explanation of the requirements are: two design matrices must be invertible linear transformations of each other, and two submatrices of the covariates which are not being tested must also be invertible linear transformations of each other. It needs to be noted that the covariates being tested are not required to be linear transformation of each other, e.g., $G_{A,R,N}$ and $G_{A,r,N}$ are not linear function of each other. Mathematically speaking, the uncertainty problem arises because four different codings of G_A have no linear transformations. When S is included in the model, as we have illustrated above, two design matrices of $(1, S, G_A)$ with different risk alleles but same XCI status become invertible linear transformations of each other. Therefore, different risk alleles result in same test statistic by applying theorem 1. This provides another reason to include sex as a covariate in the model. When sex is included, unknown risk alleles of both females and males becomes not a problem. In conclusion, we would recommend the following additive model including the sex effect at this moment:

$$Y \sim S + G_A.$$

3.2 *Sex-genotype interaction and XCI uncertainty*

For X-chromosome SNPs, genotype-sex interaction effect may exist, so that the unit effect of one copy of r or R may not be the same for males and females. The interaction is defined

by $GS = G_A \times S$. It is straightforward to check GS has two different codings depending on the risk allele of males: GS_R and GS_r as defined in Table 1.

We have explained when S is included in the model, two design matrices with different risk alleles become invertible linear transformations. Furthermore, when both S and GS are included, we can easily show all four design matrices of $(1, S, G_A, GS)$ with different risk alleles and XCI status are invertible linear transformations of each other, and for testing the null hypothesis $H_0 : \beta_A = \beta_{GS} = 0$, the design matrix of the covariates which are not being tested, i.e., $(1, S)$, remains unchanged between different coding schemes of G_A and GS. Therefore, we apply theorem 1 to show the tests with different risk alleles and XCI status are equal, and choosing correct coding of G_A and GS becomes not a issue. The relationship of invertible linear transformation between codings are summarized in Figure 2.

[Figure 2 about here.]

Figure 2 implies that in terms of testing, switching risk allele has no effect when sex is included, and the effect of inactivating X-chromosome alleles is indistinguishable to the effect of sex-genotype interaction. With S and GS included in the model, we do not need to know the risk allele and XCI status, and any one group of covariates of G_A, GS and S simply yields the same test statistic. Therefore, we now recommend including both S and GS in regression models to override the uncertainty issues about risk allele and XCI:

$$Y \sim S + G_A + GS.$$

3.3 *Dominant effects and skewed XCI*

The dominant effect G_D defined in Table 1 is invariant to risk allele and XCI status. Similar to autosome SNPs, the first reason to include the dominant effect is to capture any departure from the additive effect of the heterozygous genotype rR. For X-chromosome,

another important reason is that the dominant effect may also characterize the skewness of XCI.

Skewed XCI is the effect that one allele is more likely to be inactivated than the other for female SNPs. For homozygous genotypes rr and RR, the genotype effects are always reduced to a half because both alleles have same effect and which allele is inactivated makes no difference. For heterozygous genotype rR, if one allele is more likely to be inactivated, at population level the effect of rR will move towards to the effect of either rr or RR. For example, we denote the effect of rr by 0 and effect of RR by 1. Then rR will have an effect of either 0 or 1 at individual level depending on the inactivated allele. If two alleles are equal likely to be inactivated, at population level we expect half of r and half of R are inactivated, so that the averaged group effect of rR is $1/2$. If r is more likely to be inactivated, we expect more 1's than 0's on average, and rR has an average effect greater than $1/2$. With skewed inactivation, the effect of rR ranges from 0 to 1, so that the skewness is equivalent to a dominant effect making the the effect of rR different from $1/2$. In conclusion, including the covariates G_D not only captures real dominant effect, but also represents any skewness of inactivation.

When XCI status is unknown, it is more likely that the amount of skewness of the inactivated SNP is also unknown. In such case we recommend including the dominant covariate G_D to explain any possible skewness. Because the coding of G_D is invariant to risk allele and XCI status, including G_D in the model does not change the linear transformation relationships specified in Figure 2. When different covariates are chosen, Table 2 summarizes for each model whether it has problem with inflated type I error and unknown risk allele (Challenge C4 and C8), sex-genotype interaction and XCI uncertainty (Challenge C5 and C6), and dominant effects and skewed XCI (Challenge C7). Other covariates representing environmental effects can be freely added to each model. The ultimate model we recommend

is

$$Y \sim S + G_A + G_D + GS,$$

which resolves all X-chromosome specific challenges, as shown in Table 2.

[Table 2 about here.]

3.4 *Analytic power comparison*

It needs to be noted that all models with sex included as the covariate are valid to test the genotype effects because the type I error is correct, even if they are not capable to handle the XCI uncertainty or skewed XCI. The problem is they may have reduced test powers if the XCI status and skewness of XCI are not correctly specified. On the other hand, the full model M_4 may also not be most powerful because it has more degree of freedoms. Hence, a systematic power comparison of the most comprehensive model versus simpler models is desired. In short, we want to compare test powers of M_1 to M_4 in Table 2.

Similar to autosome SNPs, we need to compute the asymptotic non-centrality parameters of each test statistic for power comparison. After allowing for sex, dominant and interaction effect, it becomes not a issue to specify the true risk allele, XCI status and skewness. We assume HWE for female, equal sex frequency, but unequal allele frequencies for females and males (p_f and p_m). Then genotype frequencies of $[rr, rR, RR, r, R]$ are

$$[(1-p_f)^2/2, p_f(1-p_f), p_f^2/2, (1-p_m)/2, p_m/2].$$

We use the same technique as described in section 2 to define $\beta = (\beta_0, \beta_S, \beta_A, \beta_D, \beta_{GS}) = c/\sqrt{n}$ and fix the value of c so that the non-centrality parameter under alternative hypothesis converges to finite number as $n \to \infty$. Then the non-centrality parameter for the tests under model M_4 can be similarly computed following Cox and Hinkley (1974). M_1, M_2 and M_3 are misspecified models, and we need a re-parametrization of the covariates to simplify

the computation of non-centrality parameters. The technical details are provided in Web Appendix C and D.

The theoretical test power of all 4 tests are then computed from the non-centrality parameters. By comparing two chi-squared distributions of 1 and 3 degree of freedoms, we show in Web Appendix A that the maximum power loss by omitting both G_D and GS is 0.188, regardless of type I error, sample size and effect sizes. To see potential power gains by including G_D and GS, we want to compute test powers with different dominant effects and interaction effects. Because the XCI status and risk allele are unknown, we specify the averaged effect size (linear regression) or averaged log odds ratios (logistic regression) under each genotype group, i.e., $\mu_{rr}, \mu_{rR}, \mu_{RR}, \mu_r$ and μ_R, which can be estimated in practice without knowing the XCI status and risk allele. We fix $\mu_{rr} = -0.3, \mu_{RR} = 0.3$ and $\mu_r = 0$, and change μ_{rR} and μ_R from -0.6 to 0.6. Fixing μ_{rr} and μ_{RR} is equivalent to fixing the additive effect, and changing μ_{rR} is equivalent to changing the dominant effect, where $\mu_{rR} = 0$ corresponds to no dominant effect. Similarly, different μ_R represent different strengths of interaction effect. We show in Appendix A that the maximum power loss 0.188 is reached when $\alpha = 0.0008$, so we choose $\alpha = 0.0008$ to represent the worst case scenario. We keep $n = 1000$ and risk allele frequency $p_m = p_f = 0.2$ and 0.5. Results under other allele frequencies are presented in Web Appendix E. The test statistics from M_1 and M_2 depend on the coding of G_A and G_D. Without loss of generality, we use $G_{A,R,I}$ and GS_R for all 4 tests. One can easily choose the other codings of G_A and GS and repeat the power computation. Because the asymptotic non-centrality parameter under logistic model is equal to linear model with variance of error $\sigma^2 = 4$, Figure 3 represents test power comparisons under both linear and logistic model.

[Figure 3 about here.]

The 1 df model M_1 may have significant power loss compared to the full model as shown

by all panels. When allele frequency is 0.2, model M_2 may lose power dramatically when both the dominant and interaction effects are strong. When allele frequency is 0.5, model M_3 may not perform as good as the full model when dominant effects are strong. Therefore, we conclude that testing $\beta_A = \beta_D = \beta_{GS} = 0$ from the full model M_4 maintains overall best performance. Compared to the maximum power loss of 0.188, the power gain can be as much as 0.7 (e.g., when $\mu_{rR} = 0.6$ and $\mu_R = -0.6$) as shown in Figure 3. It implies the full model is not only robust to all the challenges, but also powerful for testing the additive effects along with various dominant and interaction effects. Therefore, it worths to consider including all G_A, G_D and G_{GS} as the covariates in practice when XCI status and/or skewness is unknown and the strength of dominant and interaction effects is not clear. After the power comparison, we still recommend the full model:

$$Y \sim S + G_A + G_D + GS.$$

4. Discussion

The assumption that $\beta \to 0$ under alternative hypothesis may not be quite intuitive, but it seems to be a common assumption when studying the theoretical properties of chi-squared tests (Cox and Hinkley, 1974; Begg and Lagakos, 1992, 1993; Neuhaus, 1998). Let $\beta = (\beta_1, \beta_2)$, where the null hypothesis is $\beta_1 = 0$ and β_2 is the nuisance parameter not being tested. Among the common practices, there is no doubt to adopt a sequence of alternative hypothesis of β_1 converging to 0, but it is not quite clear whether β_2 should also be assumed to converge to 0. In the context of GWAS, we believe it is most reasonable to assume β_1 and β_2 converge to 0 at similar rates, because both parameters denote the genotype effect of the same SNP, and there is no reason to believe the additive, dominant and interaction effects are on different scale for any sample size. Therefore, we assume both β_1 and β_2 converge to 0 as $n \to \infty$ at the rate of $1/\sqrt{n}$.

We have shown in X-chromosome association study, Sex should be included for correct type I error. For autosome study, sex is usually not included, but the result from X-chromosome suggests that when sex is a confounding variable, e.g., female and male allele frequencies unequal, it should also be included as a covariate for autosome analysis. If the risk allele is uncertain, we also recommend including sex to bypass the uncertainty.

When allele frequency difference is significant and females and males have different minor alleles, it may become unclear that if females and males have the same risk allele or each sex has its own risk allele. As the interaction effects being allowed in the models, we are essentially allowing for different risk alleles for females and males. Switching the risk allele for males is equivalent to adding an interaction effect. Following Theorem 1, it is also easy to check when GS is included, switching the risk allele only for males or females will not change the test statistic.

Although the full model on X-chromosome is robust to XCI uncertainty, it is not capable to determine whether XCI occurs or not. It is possible to detect XCI by biological experiments (Carrel and Willard, 2005), but statistical tests for XCI status may also be desired. For quantitative trait, Ma et al. (2015) proposed a variance-based test for detecting XCI. However, we do not find any statistical approach to testing XCI for binary trait up to date. It is a more challenging problem, because the binary outcomes can only yield a point estimate of percentage of cases, and do not have a variance structure similar to quantitative traits under each genotype group. Further studies are required to develop a statistical test under binary logistic model.

Acknowledgements

The authors would like to thank Dr. Lisa J. Strug for providing the cystic fibrosis application data, and Prof. Mike Evans for suggestions that have improved the presentation of the paper.

This research is funded by the Natural Sciences and Engineering Research Council of Canada (NSERC) to RVC and LS, and the Canadian Institutes of Health Research (CIHR) to LS.

References

Bagos, P. G. (2013). Genetic model selection in genome-wide association studies: robust methods and the use of meta-analysis. *Statistical Applications in Genetics and Molecular Biology* **12,** 285–308.

Begg, M. D. and Lagakos, S. (1992). Effects of mismodeling on tests of association based on logistic regression models. *The Annals of Statistics* **20,** 1929–1952.

Begg, M. D. and Lagakos, S. (1993). Loss in efficiency caused by omitting covariates and misspecifying exposure in logisticregression models. *Journal of the American Statistical Association* **88,** 166–170.

Bush, W. S. and Moore, J. H. (2012). Chapter 11: Genome-wide association studies. *PLoS Computational Biology* **8,** e1002822.

Carrel, L. and Willard, H. F. (2005). X-inactivation profile reveals extensive variability in X-linked gene expression in females. *Nature* **434,** 400–404.

Chen, B., Craiu, R. V., and Sun, L. (2017). Bayesian model averaging for the X-chromosome inactivation dilemma in genetic association study. *arXiv:1704.01207* .

Clayton, D. G. (2008). Testing for association on the X chromosome. *Biostatistics* **9,** 593–600.

Clayton, D. G. (2009). Sex chromosomes and genetic association studies. *Genome Medicine* **1,** 110.

Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics: Chapter 9.* London: Chapman and Hall.

Gao, F., Chang, D., Biddanda, A., Ma, L., Guo, Y., Zhou, Z., and Keinan, A. (2015). XWAS: A software toolset for genetic data analysis and association studies of the X chromosome. *Journal of Heredity* **106,** 666–671.

Hickey, P. F. and Bahlo, M. (2011). X chromosome association testing in genome wide association studies. *Genet Epidemiol* **35,** 664–670.

Hill, W. G., Goddard, M. E., and Visscher, P. M. (2008). Data and theory point to mainly additive genetic variance for complex traits. *PLOS Genet* **4,** e1000008.

Konig, I. R., Loley, C., Erdmann, J., and Ziegler, A. (2014). How to include chromosome X in your genome-wide association study. *Genet Epidemiol* **38,** 97–103.

Loley, C., Ziegler, A., and Konig, I. R. (2011). Association tests for X-chromosomal markers – a comparison of different test statistics. *Human Heredity* **71,** 23–36.

Ma, L., Hoffman, G., and Keinan, A. (2015). X-inactivation informs variance-based testing for X-linked association of a quantitative trait. *BMC Genomics* **16,** 241.

Neuhaus, J. M. (1998). Estimation efficiency with omitted covariates in generalized linear models. *Journal of the American Statistical Association* **93,** 1124–1129.

Ozbek, U., Lin, H. M., Lin, Y., Weeks, D. E., Chen, W., and et al. (2018). Statistics for x-chromosome associations. *Genetic Epidemiology* **42,** 539–550.

Sasieni, P. D. (1997). From genotypes to genes: doubling the sample size. *Biometrics* **53,** 1253–1261.

Wang, J., Talluri, R., and Shete, S. (2017). Selection of X-chromosome inactivation model. *Cancer Informatics* **16,** 1–8.

Wang, J., Yu, R., and Shete, S. (2014). X-chromosome genetic association test accounting for X-inactivation, skewed X-inactivation, and escape from X-inactivation. *Genet Epidemiol* **38,** 483–493.

Wise, A. L., Gyi, L., and Manolio, T. A. (2013). eXclusion: toward integrating the X chromosome in genome-wide association analyses. *Am J Hum Genet* **92,** 643–647.

Zheng, G., Joo, J., Zhang, C., and Geller, N. L. (2007). Testing association for markers on the X chromosome. *Genet Epidemiol* **31,** 834–843.

Zhongxue, C., Ng, H. K. T., Li, J., Liu, Q., and Huang, H. (2017). Detecting associated single-nucleotide polymorphisms on the X chromosome in case control genome-wide association studies. *Statistical Methods in Medical Research* **26,** 567–582.

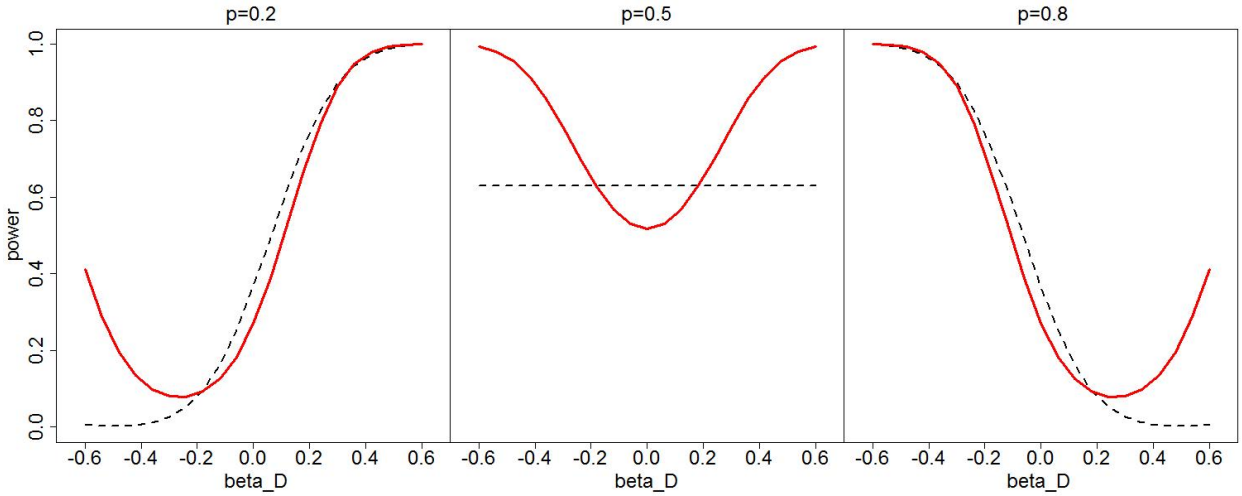


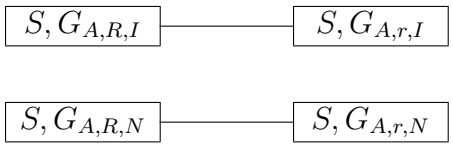
Figure 1. Power comparison of the additive model and genotypic model for autosome SNPs. Additive effect $\beta_A = 0.3$; dominant effect β_D changes from -0.6 to 0.6; allele frequency $p = 0.2, 0.5$ or 0.8 for each column. Black dash lines for testing $\beta_A = 0$ under additive model and red solid lines for testing $\beta_A = \beta_D = 0$ under genotypic model.

(a) $Y \sim \beta_0 + \beta_A G_A$
$H_0 : \beta_A = 0$

(b) $Y \sim \beta_0 + \beta_A G_A + \beta_D G_D$
$H_0 : \beta_A = 0$ or $H_0 : \beta_A = \beta_D = 0$

(a) $Y \sim \beta_0 + \beta_S S + \beta_A G_A$
$H_0 : \beta_A = 0$

(b) $Y \sim \beta_0 + \beta_S S + \beta_A G_A + \beta_D G_D$
$H_0 : \beta_A = \beta_D = 0$



(a) $Y \sim \beta_0 + \beta_S S + \beta_A G_A + \beta_{GS} GS$
$H_0 : \beta_A = \beta_{GS} = 0$

(b) $Y \sim \beta_0 + \beta_S S + \beta_A G_A + \beta_D G_D + \beta_{GS} GS$
$H_0 : \beta_A = \beta_D = \beta_{GS} = 0$

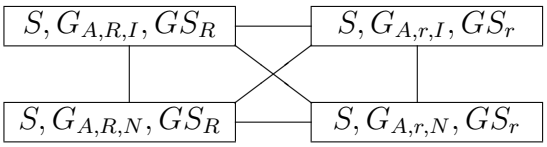


Figure 2. Linear transformations of regression models. The subscripts r or R represents risk allele, and I or N denotes X-chromosome inactivated or not inactivated. Two group of codings connected by a straight line implies an invertible linear transformation and they yield same test statistic. Part (a) corresponds to models and tests without dominant covariate G_D; part (b) corresponds to models and tests with G_D included. Whether including G_D or not has no effect to the linear relationships.

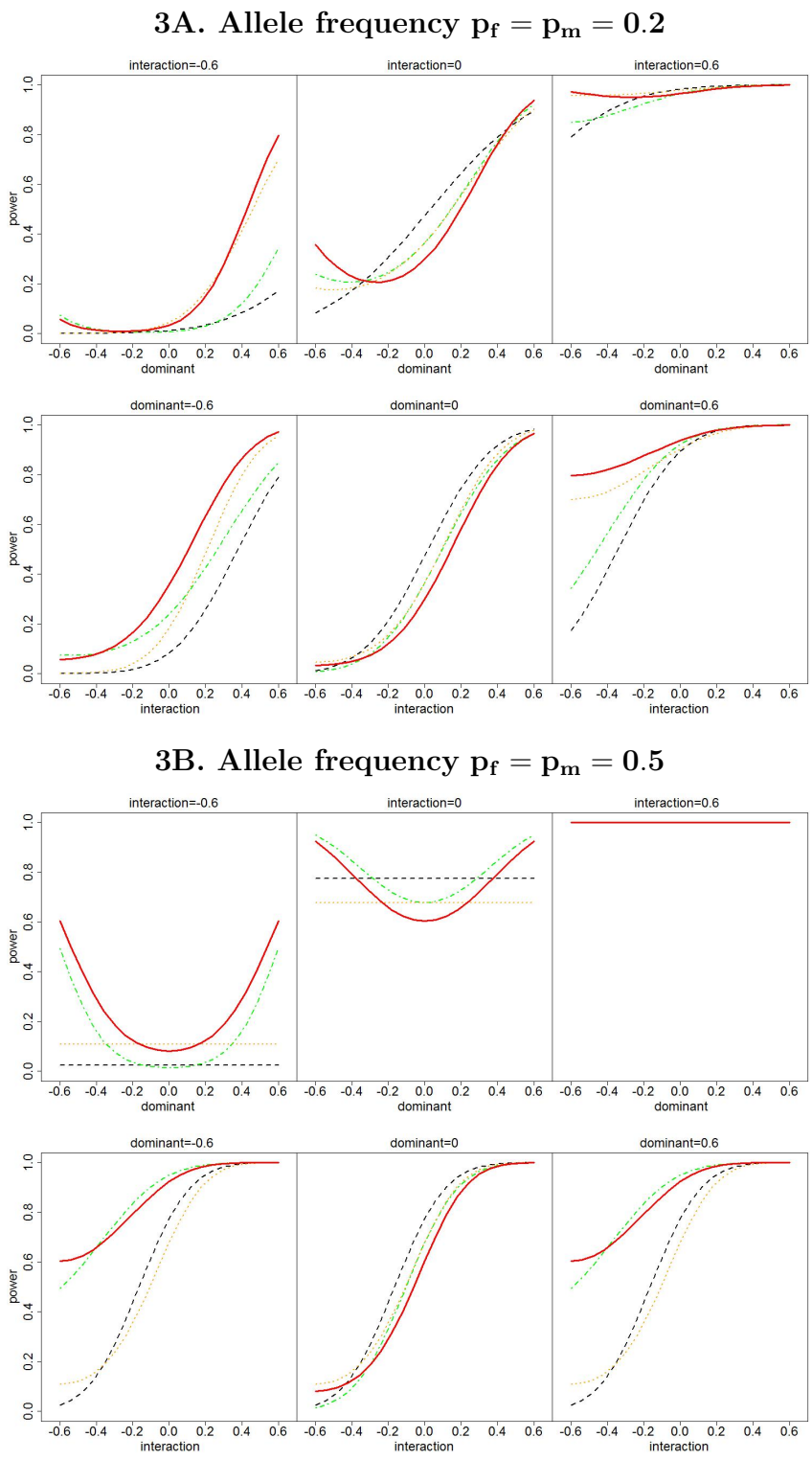


Figure 3. Power comparison for X chromosome SNPs. Black dash lines for testing $\beta_A = 0$, green dotdash lines for testing $\beta_A = \beta_D = 0$, orange dotted lines for testing $\beta_A = \beta_{GS} = 0$ and red solid lines for testing $\beta_A = \beta_D = \beta_{GS} = 0$. **Upper panels:** power vs. dominant effect (represented by the mean effect of genotype group μ_{rR}); interaction effect (represented by the mean effect of genotype group μ_R) is -0.6, 0 or 0.6 for each column. **Lower panels:** power vs. interaction effect (represented by the mean effect of genotype group μ_R); dominant effect (represented by the mean effect of genotype group μ_{rR}) is -0.6, 0 or 0.6 for each column.

Table 1

Genotype coding of the additive, dominant, sex-genotype interaction and sex effects. The interaction effects $GS = G_A \times S$. The subscripts R and r representing risk alleles, I or N denoting X-chromosome inactivated or not inactivated.

Effect	Covariate	Risk allele	X-chromosome inactivated	Genotype coding				
				rr	rR	RR	r	R
Additive G_A	$G_{A,R,I}$	R	Yes	0	0.5	1	0	1
	$G_{A,r,I}$	r	Yes	1	0.5	0	1	0
	$G_{A,R,N}$	R	No	0	1	2	0	1
	$G_{A,r,N}$	r	No	2	1	0	1	0
Dominant G_D	G_D	Either	Either	0	1	0	0	0
Interaction GS	GS_R	R	Either	0	0	0	0	1
	GS_r	r	Either	0	0	0	1	0
Sex S	S	Either	Either	0	0	0	1	1

Table 2

Candidate models having a problem to X-specific challenges or not. Challenges C4: sex confounded, C5: genotype-sex interaction, C6: X chromosome inactivation (XCI) vs. no inactivation, C7: Random vs. Skewed XCI and C8: risk alleles unknown. √ means no problem and × indicates a problem.

Model	H_0	C4/C8	C6/C7	C5
$Y \sim G_A$	$\beta_A = 0$	×	×	×
$M_1 : Y \sim S + G_A$	$\beta_A = 0$	√	×	×
$M_2 : Y \sim S + G_A + G_D$	$\beta_A = \beta_D = 0$	√	×	√
$M_3 : Y \sim S + G_A + GS$	$\beta_A = \beta_{GS} = 0$	√	√	×
$M_4 : Y \sim S + G_A + G_D + GS$	$\beta_A = \beta_D = \beta_{GS} = 0$	√	√	√