Published in final edited form as:

Genet Epidemiol. 2018 September ; 42(6): 539–550. doi:10.1002/gepi.22132.

STATISTICS FOR X-CHROMOSOME ASSOCIATIONS

Umut Özbek^{1,2,*}, Hui-Min Lin³, Yan Lin³, Daniel E. Weeks^{3,4}, Wei Chen^{3,4,5}, John R. Shaffer⁴, Shaun M. Purcell^{6,7,8,9,10}, and Eleanor Feingold^{3,4}

¹Department of Population Health Science and Policy, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA

²Tisch Cancer Institute, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA

³Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA 15261, USA

⁴Department of Human Genetics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA 15261, USA

⁵Department of Pediatrics, Children's Hospital of Pittsburgh of UPMC, Pittsburgh, PA 15224, USA

⁶Division of Psychiatric Genomics, Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA

⁷Institute for Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA

⁸Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA

⁹Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA 02114, USA

¹⁰Department of Psychiatry, Brigham & Women's Hospital, Harvard Medical School, Boston, MA 02115, USA.

Abstract

In a genome-wide association study (GWAS), association between genotype and phenotype at autosomal loci is generally tested by regression models. However, X chromosome data are often excluded from published analyses of autosomes because of the difference between males and females in number of X chromosomes. Failure to analyze X chromosome data at all is obviously less than ideal, and can lead to missed discoveries. Even when X chromosome data are included, they are often analyzed with suboptimal statistics. Several mathematically sensible statistics for X-chromosome association have been proposed. The optimality of these statistics, however, is based on very specific simple genetic models. In addition, while previous simulation studies of these statistics have been informative, they have focused on single-marker tests and have not considered the types of error that occur even under the null hypothesis when the entire X chromosome is scanned. In this study, we comprehensively tested several X chromosome association statistics using simulation studies that include the entire chromosome. We also considered a wide range of

*Correspondence to: Umut Özbek, Department of Population Health Science and Policy, Icahn School of Medicine at Mount Sinai, One Gustave L. Levy Place, Box 1077, New York, NY, 10029, USA. umut.ozbek@mountsinai.org Phone: 1 212 659-9633.

trait models for sex differences and phenotypic effects of X inactivation. We found that models that do not incorporate a sex effect can have large type I error in some cases. We also found that many of the best statistics perform well even when there are modest deviations, such as trait variance differences between the sexes or small sex differences in allele frequencies, from assumptions.

Keywords

Genetic association study; X chromosome; GWAS

INTRODUCTION

In genome wide association studies (GWAS) in humans, the first step after data cleaning is testing single nucleotide polymorphisms (SNPs) for association with a trait. Analyzing autosomal markers is more straightforward than analyzing X-chromosomal markers. Testing for association on the X chromosome, which makes up 5% of the female genome, requires specialized analysis methods – methods developed for analyzing autosomal data are not directly applicable to X chromosome data because males have only one copy of the X chromosome. Very often X chromosome data are not analyzed. For example, from January 2010 through March 2012, only 33% of the GWAS reported X chromosome results [Wise, et al. 2013]. Furthermore, from January 2017 through June 2017, only 21% of the GWAS published on Nature Genetics and PLOS Genetics reported X chromosome results. Moreover, even when the X chromosome is analyzed, often suboptimal statistics are used. A recent editorial in Nature Medicine points out that the failure to analyze X chromosomal data properly or at all now extends to the latest round of sequencing-based GWAS and reemphasizes the importance of assessing the influence of sex chromosomes and the extra effort that should be put in to include them in genomic analysis [2017].

Several promising statistical methods for X chromosome association testing have been developed [Clayton 2008; Zheng, et al. 2007]. However, there is a need for complete testing of the X chromosome methods. The best of the proposed statistics has been shown to have correct type I error and be most powerful only when there is no sex difference in allele frequencies [Clayton 2008]. Therefore, it is still not known how the statistic will behave when we scan an entire chromosome in which there is some random variation in allele frequencies between the sexes - whether it will still find the correct loci or will just pick out the ones that have the largest sex-specific allele frequency differences by chance. In this study we test several commonly-used and newly-proposed X chromosome statistics using real chromosome-wide data in order to fully understand the practical performance of the statistics. We consider the performance of the statistics under a variety of trait generating models - situations in which the male and female case-control ratios are either the same (balanced design) or different (unbalanced design), trait models with and without sex differences in the phenotype distribution, dichotomous and quantitative traits, and both rare and common minor marker alleles. We also study the behavior of statistics under various X inactivation models.

X Chromosome Test Statistics

Genotype-phenotype association is generally tested by chi-square tests, most often Armitage's trend test, or regression models. The phenotype can be dichotomous or quantitative, and genotypes can be coded in different ways. Additive coding for genotypes is typical, where at autosomal loci the genotypes are coded as (0, 1, 2). In order to test association at X chromosome loci, female genotypes are coded as (0, 1, 2) and, because males have only one X chromosome, male genotypes are coded as (0, 1) or (0, 2). In proposing X chromosome association statistics, it is necessary to consider the male genotype coding, but also male/female differences in phenotype distribution and allele frequency, and Hardy-Weinberg equilibrium (HWE) assumptions.

In this study, we evaluated six commonly-used regression models for X-chromosome association and three additional X chromosome statistics that will be described below. The regression models we included are those given in equations 1 – 6 below. P , G and S stand for phenotype, genotype and sex respectively. In the odd numbered regression models, we coded males as (0,1) and in the even numbered models, we coded males as (0,2). Note that these regression models, even when they include sex as a covariate, do not account for male-female differences in phenotypic or genotypic variance.

$$\text{Regression model G1: } P \sim G(0, 1) \quad (1)$$

$$\text{Regression model G2: } P \sim G(0, 2) \quad (2)$$

$$\text{Regression model G1S: } P \sim G(0, 1) + S \quad (3)$$

$$\text{Regression model G2S: } P \sim G(0, 2) + S \quad (4)$$

$$\text{Regression model G1xS: } P \sim G(0, 1) + S + G(0, 1) * S \quad (5)$$

$$\text{Regression model G2xS: } P \sim G(0, 2) + S + G(0, 2) * S \quad (6)$$

The statistics proposed by Clayton [2008] improve on these regression models (at least in theory) by using generalized linear model score tests based on genotype-phenotype covariance. They treat males the same as homozygote females (0,2 coding), but also account for variance differences. They do not lose power (in contrast to a stratified analysis) even if

the phenotype varies between sexes as long as allele frequency does not [Clayton 2008]. To compute the Clayton statistics, let subjects $1, \dots, F$ be female and $F+1, \dots, N$ be male. Y_i is the phenotype and A_i is the marker genotype for subject i . D_i is the heterozygosity indicator, which is 0 for homozygotes and 1 for heterozygotes. p , which is assumed to be same in males and females, is the allele frequency in the population as estimated from the data. The 2 degree-of-freedom test statistic for X chromosome data is:

$$T_2 = U^T \hat{V}^{-1} U \sim \chi^2_2 \quad (7)$$

Where

$$U = \begin{bmatrix} U_A \\ U_D \end{bmatrix} = \begin{pmatrix} \sum_{i=1}^N (Y_i - \bar{Y}) A_i \\ \sum_{i=1}^F (Y_i - \bar{Y}_F) D_i \end{pmatrix} \quad (8)$$

$$\hat{V} = \hat{V}_F \sum_{i=1}^F (Y_i - \bar{Y})^2 + \hat{V}_M \sum_{i=F+1}^N (Y_i - \bar{Y})^2 \quad (9)$$

The female and male components of the variance are:

$$\hat{V}_F = \frac{1}{F-1} \sum_{i=1}^F \begin{pmatrix} (A_i - \bar{A})^2 & (A_i - \bar{A})(D_i - \bar{D}_F) \\ (A_i - \bar{A})(D_i - \bar{D}_F) & (D_i - \bar{D}_F)^2 \end{pmatrix} \quad (10)$$

$$\hat{V}_M = \begin{pmatrix} 4p(1-p) & 0 \\ 0 & 0 \end{pmatrix} \quad (11)$$

The Clayton 1 degree-of-freedom test statistic is:

$$T_1 = U_1^2 / \hat{V}_{11} \sim \chi^2_1 \quad (12)$$

Since the Clayton 1 df statistic uses the 0, 2 male genotype coding, we will refer to this statistic as the “C2” statistic. For autosomal loci, the variance does not include the male component, $\hat{V}_M$, and $\hat{V}_F$ is calculated over all subjects. Clayton also proposed a regression generalization of C2, where phenotype is a dependent variable and sex is added as a covariate, which we will refer to as the “C2S” statistic.

Zheng et al. [2007] proposed a very different test statistic for X-chromosome association of a dichotomous trait, which is essentially a weighted average of separate male and female statistics. Their statistic (which we will refer to as the ‘Z’ statistic) is

$$Z_{mfg}^2 = \left(\sqrt{\frac{n_f}{n_m + n_f}} Z_{fG} + \sqrt{\frac{n_m}{n_m + n_f}} Z_m \right)^2 \sim \chi_1^2 \quad (13)$$

Where

$$Z_{fG} = \frac{\frac{1}{n_f^2} \left[s_f \left(\frac{1}{2} r_{f1} + r_{f2} \right) - r_f \left(\frac{1}{2} s_{f1} + s_{f2} \right) \right]}{\left[r_f s_f \left[n_f \left(\frac{1}{4} n_{f1} + n_{f2} \right) - \left(\frac{1}{2} n_{f1} + n_{f2} \right)^2 \right] \right]^{\frac{1}{2}}} \quad (14)$$

$$Z_m = \frac{\frac{1}{n_m^2} (r_m s_{m0} - s_m r_{m0})}{(n_{m0} n_{m1} r_m s_m)^{\frac{1}{2}}} \quad (15)$$

r_{mi} (r_{fi}) and s_{mi} (s_{fi}) are number of male (female) cases and controls respectively having genotype i and $r_m = r_{m0} + r_{m1}$ ($r_f = r_{f0} + r_{f1} + r_{f2}$), $s_m = s_{m0} + s_{m1}$ ($s_f = s_{f0} + s_{f1} + s_{f2}$), n_m (n_f) is number of males (females), and $n_{mi} = r_{mi} + s_{mi}$ ($n_{fi} = r_{fi} + s_{fi}$). Their statistic is based on sex-specific allele frequencies. Therefore, male genotype coding is not an issue. However, it assumes HWE in females and it does not take into account X inactivation.

Zheng et al. [2007] tested their statistic in limited simulation studies. They showed that the tests that assume HWE are more powerful than the tests that are robust to departures from HWE. Hickey and Bahlo [2011] performed more extensive simulation studies of X chromosome association testing in GWAS to investigate the effects of the sex ratios and allele frequencies in the case and control cohorts on the size and power of eight test statistics under three different disease models accounting for X-inactivation [Hickey and Bahlo 2011]. Unlike the simulations by Zheng et al., Hickey and Bahlo considered different numbers of males and females in a GWAS, a full spectrum of allele frequencies, and more than one disease prevalence. They concluded that Clayton’s test statistic is robust and powerful across a wide range of simulation parameters. In this paper, we extend this work further to consider a broader range of trait models and methods and to consider the effects of scanning an entire chromosome of data to detect the most significant loci.

MATERIALS AND METHODS

We tested the type I error and power of the six regression models, the two specialized X chromosome association test statistics by Clayton and Zheng et al. described above, and a regression generalization of Clayton's test. We considered rare and common minor alleles at the marker, dichotomous and quantitative traits, and trait model variations such as variance differences induced by X inactivation. We also considered balanced and unbalanced datasets, where male and female case-control ratios are approximately equal, or not. For this study, we defined balanced and unbalanced dataset designs as follows: if the ratio of $R_f = r_f/s_f$ (number of female cases/number of female controls) to $R_m = r_m/s_m$ (number of male cases/number of male controls) is between 0.80 and 1.20 then the dataset is defined as balanced, otherwise unbalanced.

The simulations for type I error rates are based on real data for the entire X chromosome excluding the pseudo-autosomal regions. The simulations for power are based on simulated SNPs using the real data sample sizes. Our rationale for this approach is as follows. Conventional statistical theory measures the optimality of statistics in terms of having correct type I error and maximal power for a single test. But in genomic applications, we apply a test thousands to millions of times and pick out the most significant loci for further study. In that situation, it is not the expected value of the behavior of the statistic that matters, but rather the behavior of the extreme values (order statistics). For example, Clayton's test statistic was shown by Hickey and Bahlo [2011] to be most powerful, but only when there is no sex difference in allele frequencies. How will that statistic behave when we scan an entire chromosome in which there is a small amount of random variation in allele frequencies between the sexes? We performed our simulation study using real chromosome-wide data in order to fully understand the practical performance of the statistics in realistic situations.

Genotype data for our simulations came from the Gene Environment Association Studies (GENEVA) preterm birth dataset (dbGaP accession number: phs000103.v1.p1). In this GWAS dataset, there are approximately 2,000 mother-baby pairs genotyped using the Illumina Human 660W-Quad chip. We dropped mothers' data and used only 1,795 babies in our study. There are 863 female babies (393 cases and 470 controls) and 932 male babies (451 cases and 481 controls) in the dataset. PLINK was used to obtain the minor allele frequencies (MAFs) and the HWE p-values [Purcell, et al. 2007]. SNPs with HWE p-value < 0.0001 were excluded. After filtering, 12,242 X chromosome SNPs were retained for the common allele analyses (MAF>0.02) and 622 SNPs for rare allele analyses (0<MAF<0.02).

For type I error studies, we used the above-described genotype data, and simulated phenotypes (both dichotomous and quantitative) independent of all genotypes. The type I error rate was computed by simulating one set of phenotypes, and then tallying the number of positive SNPs across all the SNPs chromosome-wide. The full dataset includes 393 female cases, 470 female controls, 451 male cases, and 481 male controls, which is balanced according to our definition. To create other balanced (male and female case-control ratios are the same) and unbalanced datasets for the dichotomous traits, we dropped subsets of

males and females (who were randomly selected) to arrive at the desired ratios. We also simulated “spiked in” genotypes for 120 additional common loci with larger-than-normal absolute sex differences in allele frequency in the range of (0.07, 0.15). The minor allele frequencies of the spiked-in genotypes were in the range of (0.03, 0.50). Additional details regarding the spiked-in SNPs are provided in Table S1. Details regarding the three balanced and the six unbalanced dataset designs we created can be found in the first two columns of Table S2. Computing the error rate for only real SNPs, only spiked-in SNPs, and then both real and spiked in SNPs together allowed us to test the behavior of the statistics both for normal variation between male and female allele frequencies and for extreme situations. Figure 1 shows the density plot of the real and spiked-in SNPs’ allele frequency differences used in the type I error rate analyses.

For power analyses, we simulated 200 replicates of a single SNP-phenotype pairs (for both balanced and unbalanced designs). For dichotomous traits, we simulated two phenotype groups: cases and controls. We considered two different allele frequency assumptions for common alleles. First, given the number of cases and controls, we assumed marker allele frequencies of 0.50 and 0.46 for controls and cases respectively (same in males and females). Note that this does not imply a particular genetic risk model in females, but does assume that the type of effect is relatively similar between males and females. Then we tested the more unexpected situation in which female minor allele frequencies were 0.47 and 0.49 for controls and cases, and male allele frequencies were 0.50 and 0.46 for controls and cases (i.e., opposite effect directions in the two sexes). For the rare allele analysis, we tested allele frequencies of 0.02 and 0.01 for controls and cases (same in males and females); and the scenario with female frequencies of 0.025 and 0.015 for controls and cases, and male allele frequencies of 0.02 and 0.01 for controls and cases. These scenarios are not comprehensive, but were chosen to examine the behavior of the statistics when male and female allele frequencies are similar and when they are different. To explore the behavior of the association statistics when applied to quantitative phenotypes, we used the genetic models as shown in Table 1. We simulated phenotypes from three different phenotype distributions for each sex. Power was calculated as the fraction of positive tests across all 200 replicates.

Test Statistics

After simulating the data, we analyzed it using three approaches: regression analysis, Clayton’s, and Zheng’s methods. For regression analysis, we fitted the models (G1-G2xS) introduced in equations 1–6. Then we applied Clayton’s method using Clayton’s 1-df test statistic (C2) as shown in equation 12 and a regression generalization of Clayton’s test where sex was added as a covariate (the C2S statistic). Finally, we applied Zheng’s test as shown in equation 13 (the Z statistic).

RESULTS

Dichotomous Trait Analyses for Common Alleles

Table S2 shows complete results for type I error for dichotomous traits. When the dataset is balanced, we observed that all type I error rates fall in the Bradley’s liberal criterion range of

0.025 to 0.075 [Bradley 1978] and none of them exceeds 0.058. In the unbalanced designs regression models G1 and G2 and the Clayton 1-df statistic C2 had extremely high type I error for the spiked-in SNPs (see results for the “Spiked-in” dataset in Table S2), and regression model G1 had extremely high type I error for the real SNPs. If the datasets are very unbalanced (e.g. datasets U_0.11 and U_8.40), the type I error may be as high as 0.89 if a sex covariate is not included. As detailed in the Appendix, the reason for this is that, under the null hypothesis, in a case control study cohort, the overall disease probability, ignoring sex, for a specific genotype group is a function of the conditional disease probabilities given sex and the minor allele frequency. The disease probabilities within the samples are equal across different genotype groups when the proportion of cases is the same in both sexes (i.e. balanced design). The probability of disease for different genotype groups can differ dramatically when the proportion of cases differs in males and females (Figure S1). Figure 2 shows the type I error rates for the real SNPs for the tests of dichotomous traits, excluding model G1 so that the other analysis models can be compared more appropriately. We also provide the Q-Q plots for the most balanced set (B_0.89) and one of the most unbalanced sets (U_8.40) (Figures S2-3). For the balanced design (Figure S2) all p-values are consistent with the theoretical distribution. However, for the unbalanced design, analysis model G1 drastically deviates from the theoretical distribution while the rest of the analysis models are consistent as expected (Figure S3).

To further our goal of understanding how the statistics perform under realistic study conditions, we asked whether a “top 10” gene list might be dominated by aberrant SNPs (such as our spiked in SNPs with large sex differences in allele frequency) even if the statistic behaves well in most cases and even under the null hypothesis. Table 2 shows the number of SNPs in the “top 10” list (10 smallest p-values) that were spiked in. For the balanced design, we again see no problem, but for the unbalanced design the top 10 list is indeed dominated by the spiked in SNPs for the same three methods that showed problems in Table S2 – regression models G1 and G2, and the Clayton C2 statistic.

In the dichotomous phenotype power analysis (Figure 3, Table S3), we observed that under design B_0.89 when the male and female allele frequencies are equal in cases and controls, regression model G1S has highest power among the statistics that had robust type I error across all designs. However, when the male and female allele frequencies are unequal, regression model G2S and the Clayton C2S statistic are most powerful under design B_0.89. In unbalanced designs, among the statistics with robust type I error, G1S and Z have the highest power for the model with similar male and female effects, though G2S and C2S are better in some cases for the model with opposite sex effects. Note that a power estimate of 0.5 has a standard error of approximately 0.036.

Quantitative Trait Analyses for Common Alleles

Type I error results for quantitative phenotypes are given in Table S4. Results are very similar to those for dichotomous phenotypes. For regression model G1, we observed very high type I error rates when male and female phenotype means were different (essentially equivalent to an unbalanced design). However, regression models with male genotypes coded as (0,2), and/or the models with a sex covariate have well-controlled type I error rates.

In quantitative phenotype power analysis (Table S5), among the methods that have well-controlled type I error rates, regression model G2S and Clayton's C2S statistic again have highest power when males with the B genotype have the same mean as homozygous BB females. However, when the B males' mean is the same as heterozygous AB females, G1S is more powerful.

Rare Allele Analysis

Our rare allele analyses focused on SNPs with $MAF < 0.02$. Since these SNPs yield tests for which asymptotic assumptions are less likely to hold, we wanted to study the behavior of the statistics on this subset of SNPs separately. On the X chromosome there are 622 SNPs with $0 < MAF < 0.02$ and HWE p -value > 0.0001 .

Figure 4 shows the type I error rates of the X chromosome statistics for rare minor alleles for dichotomous phenotypes. Although the type I error rates of regression models G1xS and G2xS are a little higher than the others, all rates fall within Bradley's liberal criterion range of 0.025 to 0.075 [Bradley 1978]. A more detailed breakdown is provided in Table S6. Most of the results are consistent with those for the common SNPs, except that the type I error for regression model G1 does not seem to be severely inflated as we observed in the common SNPs. To further investigate this, we plotted the type I error rate by MAF in our data and observed that the type I error of regression model G1 increases with the MAF for the unbalanced design U_0.34 (Figure 5 (b)) and the type I error rate increases faster for the unbalanced design U_8.40 (Figure 5 (c)). The type I error is well controlled for the balanced design B_0.89 (Figure 5 (a)). Similar results were observed for quantitative phenotypes.

To explain this phenomenon, we illustrate the case of the quantitative phenotype in Figure 6. Under the null, for the unbalanced design U_0.34 (assuming the mean of the females is higher than that of the males), an apparent positive association is detected for a common allele (Figure 6 (a)). However, for a rare allele, the homozygous minor allele group has very few or no data points; thus the positive association is not significant, thus not affecting the type I error as much (Figure 6 (b)).

Figure 7 shows the results of power analyses for the dichotomous phenotype rare allele scenario with minor allele frequencies of 0.02 and 0.01 for controls and cases (same in males and females). We observed that the regression model G1S and regression model G1xS have relatively higher power in the most of the sampling designs. However, the power of regression model G1 is very unstable across different sampling designs. Similar results were observed for a quantitative phenotype.

X Inactivation

Another issue that should be considered in the evaluation of X chromosome association statistics is X inactivation. X inactivation is transcriptional silencing of the majority of one of the X chromosomes in a complex manner in females [Lyon 1961]. Because of X inactivation, none of the statistics described above are actually based on a correct trait model. The X inactivation model affects the presumed mean and variance for the female heterozygote genotype group, and thus in theory could affect the choice of the best statistic.

There has been some work on statistics that model X inactivation [Clayton 2009; Wang, et al. 2014]. For example, Clayton suggests that for allelic tests, each allele in females should be counted as $\frac{1}{2}$ to reflect the dosage compensation for X inactivation [Clayton 2009].

We propose a more realistic trait model, in which the allele silenced is not uniform throughout the organism - it may or may not even be uniform in a particular tissue. We suggest that a heterozygote female could randomly be anything between pure A and pure B. If this is the case, the variance and the mean of a quantitative trait for female heterozygotes would not be simple as in the case above. In particular, the variance would be much higher than the one we could estimate from the sample. In quantitative traits, most X-inactivation models result in higher trait variance for heterozygote females than for homozygote males or females. In theory, this variance difference could increase the type I error of traditional regression-based tests.

We studied the behavior of the statistics under different X-inactivation models. To test the potential effects of different variances in males and females, we simulated quantitative phenotype variables with uniform and genotype-specific variances. We compared Clayton's statistic and regression methods in terms of type I error and power under X-inactivation models, where randomly one of the X chromosomes in females is transcriptionally inactivated, by using simulated datasets. From these experiments, we observed that type I error rates and powers of the robust linear regression analyses using an M estimator [Venables, et al. 2002] and Clayton's test (Clayton 1-df) are close (Table 3). Wang et al. [2014] proposed an approach of maximizing likelihood ratio of all biological possibilities of X inactivation process and showed that their approach had higher power than Clayton's test. We did not include the Wang et al. method in our comparisons because it is computationally intensive and therefore not a chromosome-scanning method; our focus is on the effects of chromosome-wide analysis on statistical performance.

DISCUSSION

Failure to analyze X chromosome data at all is obviously less than ideal, and can lead to missed discoveries – for example, the first step in the SNP quality control process in a GWAS for diabetic nephropathy was to remove the X chromosome [Pezzolesi, et al. 2009], but when the dataset was submitted to the database of genotypes and phenotypes (dbGAP), the standard pre-compute analysis by dbGAP discovered that the X-linked SNP rs16997315 was strongly associated with a p-value of 4.7×10^{-11} (according to the Phenotype-Genotype Integrator website from NCBI). Even if the X chromosome data are analyzed, suboptimal statistics may be used. To analyze X chromosome data, specialized analysis methods are needed. Although there are some statistics developed for X chromosome analysis, they assume relatively simple genetic models. Moreover, these statistics are seldom used for real data analysis, at least partly because their statistical properties (strengths and weaknesses) are not well understood.

In this study we aimed to extensively evaluate three specialized X chromosome association test statistics [Clayton 2008; Zheng, et al. 2007], C2, C2S, and Z, and compare them with regression models using realistic simulated datasets under various genetic models. We

considered balanced and unbalanced datasets; if the ratio of $R_f = (\text{number of female cases} / \text{number of female controls})$ to $R_m = (\text{number of male cases} / \text{number of male controls})$ is between 0.80 and 1.20 then the dataset is defined as balanced, otherwise unbalanced. We emphasized the behavior of the statistics, especially under the null hypothesis, when scanning the whole X chromosome, so that, for example, there may be natural variation in the difference between male and female allele frequencies. In evaluating power, we also looked for statistics with power that is robust to modeling assumptions.

We found that when the sampling design is balanced (for a dichotomous trait) or the male and female trait means are similar (for a quantitative trait), all statistics have correct type I error. However, statistics without a sex effect in the model can have extremely high type I error when the sampling is unbalanced or quantitative trait means differ between the sexes. This problem is substantially alleviated when the male genotype is coded as 0/2 instead of 0/1, except for the most extreme circumstances (such as our spiked in SNPs). Our chromosome-wide real data experiment makes it difficult to determine why the 0/2 coding works better, but does give strong confidence that the result is realistic. For unbalanced designs, regression model G1 has very high Type I error rates and should not be used. For unbalanced designs with extreme sex-specific allele frequency differences (e.g., the spiked-in SNPs), regression models G1 and G2 and the Clayton C2 statistics are not appropriate, although such large sex differences in allele frequency are quite rare in real datasets. Loley et al [2011] arrived at the same conclusion regarding the Clayton C2 statistic. They suggested checking for allele frequency differences before applying such tests, but that advice does not provide a roadmap for genome-wide application. Our recommendation is to apply a statistic that has appropriate type I error regardless of allele frequency differences between the sexes. Thus, we recommend always including a sex effect in the analysis model as a general precaution.

Our power studies did not produce a clear preference for 0/1 coding or 0/2 coding, and since any real trait is likely to have effects from multiple SNPs that may behave according to different models, we are unable to recommend one coding over the other. However, it should be reassuring to analysts to see that the power difference between 0/1 and 0/2 coding is rarely large, so an arbitrary choice should not be harmful.

For X inactivation models, although Clayton's test performs similarly to robust regression methods, both seem to have conservative type I error. Further investigation of the tests under X inactivation models may be needed.

One issue we did not address involves methods for sequenced X-chromosome data. If an allele of an X-chromosome SNP is very rare, in females we would expect to see mostly AA individuals with a few ABs and maybe one or two BBs. The (0,1,2) coding analysis model is not appropriate for this type of data because BB individuals will be influential points in the regression and can bias the analysis (Figure 6). Therefore, females should be coded as (0,1), where AB and BB females are 1. However, for males, it is not obvious whether (0,1) or (0,2) coding is appropriate.

In conclusion, this simulation study showed that available statistical tests can appropriately handle X chromosome data in genomic studies and are fairly robust to deviations from assumptions, as long as a sex effect is included in the model.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

The work was funded by the NIH grants R03 DE021425 (E.F.) and R01 HG007358 (W.C.).

APPENDIX

Disease Probabilities

Assume minor allele frequency is p , and female and male probabilities are $P(F) = P(M) = 1/2$.

Let G_0 , G_1 and G_2 be female homozygous, heterozygous and homozygous genotypes respectively and G_0 and G_1 be male genotypes.

For females we have:

$$P(F, G_0) = P(F)P(G_0|F) = \frac{1}{2}(1-p)^2$$

$$P(F, G_1) = P(F)P(G_1|F) = p(1-p)$$

$$P(F, G_2) = P(F)P(G_2|F) = \frac{1}{2}p^2$$

For males we have:

$$P(M, G_0) = P(M)P(G_0|M) = \frac{1}{2}(1-p)$$

$$P(M, G_1) = P(M)P(G_1|M) = \frac{1}{2}p$$

Under the null hypothesis, the conditional disease probabilities given sexes are:

$$P(D|F, G_0) = P(D|F, G_1) = P(D|F, G_2) = f_1$$

$$P(D|M, G_0) = P(D|M, G_1) = f_2$$

The marginal disease probabilities given genotype are:

$$P(D|G_0) = \frac{P(D, G_0)}{P(G_0)} = \frac{P(D, F, G_0) + P(D, M, G_0)}{P(G_0)} = \frac{f_1 \frac{1}{2}(1-p)^2 + f_2 \frac{1}{2}(1-p)}{\frac{1}{2}(1-p)^2 + \frac{1}{2}(1-p)} = \frac{f_1(1-p) + f_2}{(1-p) + 1}$$

$$P(D|G_1) = \frac{P(D, G_1)}{P(G_1)} = \frac{P(D, F, G_1) + P(D, M, G_1)}{P(G_1)} = \frac{f_1 p(1-p) + f_2 \frac{1}{2}p}{p(1-p) + \frac{1}{2}p} = \frac{f_1(1-p) + f_2/2}{(1-p) + 1/2}$$

$$P(D|G_2) = \frac{P(D, G_2)}{P(G_2)} = \frac{P(D, F, G_2) + P(D, M, G_2)}{P(G_2)} = \frac{f_1 \frac{1}{2}p^2}{\frac{1}{2}p^2} = f_1$$

The marginal disease probabilities are equal when $f_1 = f_2$.

REFERENCES

- 2017 Accounting for sex in the genome. *Nat Med* 23(11):1243. [PubMed: 29117171]
- Bradley JV. 1978 Robustness? *British Journal of Mathematical and Statistical Psychology* 31(2):144–152.
- Clayton D 2008 Testing for association on the X chromosome. *Biostatistics* 9(4):593–600. [PubMed: 18441336]
- Clayton DG. 2009 Sex chromosomes and genetic association studies. *Genome medicine* 1(11):110. [PubMed: 19939292]
- Hickey PF, Bahlo M. 2011 X chromosome association testing in genome wide association studies. *Genetic epidemiology* 35(7):664–70. [PubMed: 21818774]
- Lyon MF. 1961 Gene action in the X-chromosome of the mouse (*Mus musculus* L.). *Nature* 190:372–3. [PubMed: 13764598]
- Pezzolesi MG, Poznik GD, Mychaleckyj JC, Paterson AD, Barati MT, Klein JB, Ng DP, Placha G, Canani LH, Bochenski J and others. 2009 Genome-wide association scan for diabetic nephropathy susceptibility genes in type 1 diabetes. *Diabetes* 58(6):1403–10. [PubMed: 19252134]
- Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ and others. 2007 PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics* 81(3):559–75. [PubMed: 17701901]
- Venables WN, Ripley BD, Venables WN. 2002 *Modern applied statistics with S*. New York: Springer.
- Wang J, Yu R, Shete S. 2014 X-chromosome genetic association test accounting for X-inactivation, skewed X-inactivation, and escape from X-inactivation. *Genet Epidemiol* 38(6):483–93. [PubMed: 25043884]
- Wise AL, Gyi L, Manolio TA. 2013 eXclusion: toward integrating the X chromosome in genome-wide association analyses. *Am J Hum Genet* 92(5):643–7. [PubMed: 23643377]
- Zheng G, Joo J, Zhang C, Geller NL. 2007 Testing association for markers on the X chromosome. *Genet Epidemiol* 31(8):834–43. [PubMed: 17549761]

Figure 1. Allele Frequency Differences Between Females and Males.

The solid line refers to 12,242 real X chromosome SNPs with MAF > 0.02. The dashed line refers to 120 spiked-in SNPs with large allele frequency differences between females (F) and males (M).

Figure 2. Type I Error Rates of the Methods for Dichotomous Phenotypes and 12,242 Real SNPs. The red line indicates the nominal Type I error rate of 0.05, while the black lines indicate the boundaries of the Bradley's liberal criterion range. Simulations designated B and U represent balanced and unbalanced designs; the R_f/R_m ratio ranges from 0.87 to 8.40 is indicated. Details of the various simulation designs can be found in Table S2. The G1 statistic was excluded from this graph because of its high Type I error rates.

Figure 3. Power of the Methods for Dichotomous Phenotypes.

Power of the methods that had robust type I error rates across all designs where the male and female allele frequencies are equal in cases and controls, with dichotomous phenotypes. Simulations designated B and U represent balanced and unbalanced designs; the R_f/R_m ratio ranges from 0.87 to 8.40 is indicated. Details of the various simulation designs can be found in top half of Table S3. The G1 statistic was excluded from this graph because of its high Type I error rates, hence inflated power estimates.

Figure 4. Type I Error Rates of the Methods for Dichotomous Phenotypes in Rare Allele Scenarios.

The minor allele frequencies in cases and controls are 0.01 and 0.02 in both males and females. The red line indicates the nominal Type I error rate of 0.05, while the black lines indicate the boundaries of the Bradley's liberal criterion range. Simulations designated B and U represent balanced and unbalanced designs; the Rf/Rm ratio ranges from 0.87 to 8.40 is indicated. Details of the various simulation designs can be found in the first section of Table S6. The G1 statistic was excluded from this graph because of its high Type I error rates.

Figure 5. G1 Model Type I Error Rates.

G1 model ($P \sim G(0, 1)$) type I error rates over different MAF for the designs (a) B_0.89, (b) U_0.34, and (c) U_8.40.

Figure 6. Genetic Association for Unbalanced Design.

Genetic association for the unbalanced design U_0.34 under null hypothesis on the (a) common allele and (b) rare allele. Black dots and yellow dots indicate mean value of male phenotype and female phenotype, respectively. The size of dots is proportional to sample sizes within the categories. Green symbols indicate overall mean value of phenotype. Blue lines indicate fitted lines that are estimated from the regression model.

Figure 7. Power of the Methods for Dichotomous Phenotypes in Rare Allele Scenarios. The minor allele frequencies in cases and controls are 0.01 and 0.02 in both males and females. Simulations designated B and U represent balanced and unbalanced designs; the Rf/Rm ratio ranges from 0.87 to 8.40 is indicated. Details of the various simulation designs can be found in the first section of Table S7.

Table 1.

Quantitative Phenotype Power Analysis Models.

Phenotype distributions Normal (Mean, SD=13)					Common allele		Rare allele	
Mean value for male genotype		Mean value for female genotype			Male minor allele frequency	Female minor allele frequency	Male minor allele frequency	Female minor allele frequency
A	B	AA	AB	BB				
15	16	15	16	17	0.30	0.30	0.01	0.01
15	17							
15	16	15	16	17	0.33	0.30	0.015	0.01
15	17							
15	16	15	16	17	0.37	0.30	0.02	0.01
15	17							
15	16	15	16	17	0.30	0.40	0.01	0.015
15	17							
15	16	15	16	17	0.45	0.30	0.01	0.02
15	17							

Table 2.
Number of Spiked-in Snps on the Top 10 List.

Details of the simulation designs can be found in Table S2.

Design	Females (Case/Cont)	Males (Case/Cont)	G1: P~ G(0,1)	G2: P~ G(0,2)	G1S: P~ G(0,1)+S	G2S: P~ G(0,2)+S	C2: Clayton (1-df)	Z: Zheng
B_0.89	393/470	451/481	0	0	0	1	0	0
U_8.40	393/150	150/481	5	10	0	0	10	0

Table 3.
Clayton And Regression Results For Continuous Phenotypes.

Clayton and regression results based on simulated 1,000 200-sample datasets. G takes values (0,1,2) for female genotypes (AA, AB, BB) and (0,2) for male genotypes (A,B).

<i>Phenotype distribution</i>		<i>Test</i>			
<i>Heterozygote females</i>	<i>Homozygote females and males</i>	<i>Probability</i>	<i>Clayton (1-df)</i>	<i>Robust regression using M estimator</i>	<i>Linear regression</i>
N(11, 30)	N(10+G, 8)	<i>Power</i>	0.969	0.987	0.957
N(10, 30)	N(10, 8)	<i>Type I error (at 0.05 level)</i>	0.021	0.022	0.012