Master’s Thesis of Heejin Jin

Efficient statistical method for association analysis of x-linked variants

X염색체 변이의 연관성분석을 위한 효율적인 통계분석방법

August 2016

Graduate School of Public Health
Seoul National University
Public Health Major

Heejin Jin

Efficient statistical method for association analysis of x-linked variants

Sungho won

Submitting a master's thesis of Public Administration

May 2016

Graduate School of Public Health
Seoul National University
Public Health Major

Heejin Jin

Confirming the master's thesis written by
Heejin Jin
July 2016

Chair ______Ho kim_______(Seal)
Vice Chair _____Joohon Sung____(Seal)
Examiner _____Sungho Won____(Seal)

Abstract

Unlike the gene-poor Y-chromosome, the X-chromosome contains over 1,000 genes that are essential for proper development and cell viability. Thus, amount of females' X-linked gene expression would be expected to be twice as males'. To adjust this unbalancedness, one of two X-linked genes in female is often inactivated and this is known as X-chromosome inactivation (XCI). However, recent studies found that a gene for inactivation can be nonrandomly selected from two X-linked genes and XCI was not observed in some X-linked genes. This biologically complicated process has prevented the efficient statistical association analyses, and it may partially explain the relatively small finding of significantly associated X-linked variants.

Here, I propose a new statistical method robust against the uncertain biological process. The proposed method consists of two steps. First, p-values for various biological processes are calculated and combined into a single p-value with modified Fisher's method and minimum P-value. Our simulation results show that the scenario for specific XCI process with true underlying coding is the most powerful but the proposed methods are the second best.

Keyword : X-chromosome inactivation, X-linked variants, X-chromosome association analysis

Student Number : 2014-23378

Table of Contents

List of Tables

List of Figures

Introduction

Unlike the gene-poor Y-chromosome, the X-chromosome contains over 1,000 genes that are essential for proper development and cell viability. Thus, amount of females’ X-linked gene expression would be expected to be twice as males’. For equivalent amount of a gene expression between males and females, the X-chromosome inactivation (XCI) on female X-chromosome loci prevents female from having twice as many gene expression as male. However, recent studies found that genes for inactivation can be nonrandomly selected from two X-linked genes and XCI was not observed in some X-linked genes [Amos-Landgraf et al., 2006; Belmont, 1996, Busque et al., 2009; Chagnon et al., 2005; Minks et al., 2008; Plenge et al,, 2002; Struewing et al., 2006; Willard, 2000; Wong et al., 2011]. This biologically complicated process has prevented the efficient statistical association analyses, and it may partially explain the relatively small finding of significantly associated X-linked variants.

Multiple approaches for X-linked variants were proposed and firstly Clayton [Clayton, 2008] suggested two chi-squared tests with 1 and 2 degree-of-freedom tests. He assumed that effect of males’ homogeneous genotypes on phenotypes is equivalent to females’ homozygous genotypes, and females’ genotypes are coded as 0, 1, or 2, and males’ genotypes are coded as 0 or 2. The proposed method remains valid when phenotype varies between sexes, provided the allele frequency does not, and avoids the loss of power resulting from stratification by sex in such circumstances [Clayton, 2008]. This method was the most powerful when the true underlying biological model was random XCI, but it lost some power when the true underlying biological models were nonrandom XCI or XCI was not observed in some X-linked genes [Jian Wang et al. 2014]. Jian Wang et al. suggests new statistical approach for various XCI process, i.e, random XCI, nonrandom XCI or escaped XCI (XCI do not occur). They coded 0, or 2 for males' genotypes and 0, d, 2 for females' genotypes. d is for heterogeneous genotypes and can be any real number between 0 and 2 unlike the Clayton's approach. d is related with the level of skewness in the heterozygous females [Jian Wang et al. 2014], d = 1 means a random XCI which same as Clayton's additive generic model. If d is between 0 and 1, a nonrandom XCI toward the normal allele is assumed and if it is between 1 and 2, a nonrandom XCI toward the deleterious allele is expected. Males' coding depends on females' coding like that if there was random XCI or nonrandom XCI, males' coding for X is {0, 2} but in case of not occur XCI in female then {0, 1} to equalize a gene expression. However, this approach has higher power when XCI is nonrandom and is less efficient if XCI was random and XCI does not occur [Jian Wang et al. 2014]. In this thesis, I suggest new statistical methods robust against the various XCI process and extensive simulation studies showed that the proposed method preserves reasonable statistical power for all XCI models even though it is not always best.

Methods

Notations and the disease model

We assumed that there are N_m males and N_f females, and total sample size is N. We consider only X-linked variants, and there is a single allele for males. Depending on the X-chromosome inactivation process, we assume that males' genotypes are coded by 0/1, or 0/2. The former is for escaped X-chromosome inactivation process and the latter for the X-chromosome inactivation. If we denote the disease and normal alleles by A and a respectively, aa, Aa, and AA are coded by 0, d, and 2 respectively. The choice of d denotes the level of skewness for the heterozygous genotypes of females. If d is less than 1, it assumes that disease alleles of heterozygous genotypes tend to be less activated compared to normal allele, and if d is larger than 1, disease alleles tend to be more expressed. Note that d=1 indicates that randomly selected alleles for heterozygous genotypes are active. Therefore, d can have a value from 0 to 2 which depends on X-chromosome inactivation process in female. The data available from a case-control study are showed in Table 1 and 2, where A is a high deleterious allele and a is the normal allele for disease. Also, a number in parentheses means that the X score for genotype. We will use a set of scores which must be assigned to genotypes for Cochran Armitage test as weight. Also, R and S are sample size of case and control respectively, and the total sample size was denoted by N_m for male ; N_f for female. Since this is for case-control study, a number of cases and controls are fixed.

Table 1. Genotype Table (Male)

	Genotype (Male)		Total
	A (1 or 2)	a (0)	
Cases	r_1	r_0	R
Controls	s_1	s_0	S
Total	n_1	n_0	N_m

		Genotype (Female)		Total
	AA (2)	Aa (d)	aa (1)	
Cases	r_2	r_1	r_0	R
Controls	s_2	s_1	s_0	S
Total	n_2	n_1	n_0	N_f

Table 2. Genotype Table (Female)

Cochran-Armitage Trend Test for the genotype table

We assume that $(r_0,\ r_1)$ follows a binomial distribution with probabilities for genotype a and A corresponding to p_0 and p_1, and $(s_0,\ s_1)$ follows a binomial distribution with probabilities q_0 and q_1; for females, $(r_0,\ r_1,\ r_2)$ follows a trinomial distribution with probabilities for genotype aa, Aa, and AA corresponding to $p_0,\ p_1$ and p_2 and $(s_0,\ s_1,\ s_2)$ follows a trinomial distribution with probabilities $q_0,\ q_1$ and q_2. The population genotype probabilities will be expressed by $g_0,\ g_1$ and g_2 which means probability of aa, Aa and AA, respectively and K, the disease prevalence, can be written as

$$\mathrm{K} = \sum_i f_i g_i.$$

Assume that the penetrances of aa, Aa, AA as f_i, i equals to 0, 1, 2 for females and 0, 1 for males. In the above notation, p_i and q_i can be expressed as

$$p_i = \frac{f_i g_i}{K} \text{ and } q_i = \frac{(1-f_i)g_i}{1-K}.$$

Therefore, we expressed the null hypothesis as $H_0\colon p_i = q_i$ (for male i=0,1; for female i=0,1,2) [Freidlin B et al. 2002]. We worked with the difference the values in the column so we first standardized the rows to have the same sums. We choose a set of scores x_1 to x_k and form the test statistic

$$\mathrm{U} = \sum_{i=0}^{k} x_i\,(Sr_i - Rs_i)$$

Under H_0, $\mathrm{P}(\text{ Case} \mid \text{ith genotype }) = \mathrm{P}\ (\text{Control} \mid \text{ith genotype}) = \mathrm{n_i/N}$ and $\mathrm{E(U)} = 0$. Returning to generic scores x_i, we calculated the variance of U as:

$$\mathrm{var(U)} = var\left(\sum_{i=0}^{k} x_i(Sr_i - Rs_i) \right)$$

$$= \mathrm{var}\left(\mathrm{S}\sum_{i=0}^{k} x_i r_i - R\sum_{i=0}^{k} x_i s_i \right)$$

$$= S^2 var\left(\sum_{i=0}^{k} x_i r_i \right) + R^2 var\left(\sum_{i=0}^{k} x_i s_i \right)$$

$$= S^2 \left[\sum_{i=0}^{k} {x_i}^2 var(r_i) + 2\sum_{i=0}^{k-1} \sum_{j=i+1}^{k} x_i x_j cov(r_i, r_j) \right]$$

$$+ R^2 \left[\sum_{i=0}^{k} {x_i}^2 var(s_i) + 2\sum_{i=0}^{k-1} \sum_{j=i+1}^{k} x_i x_j cov(s_i, s_j) \right].$$

(Under H_0)

$$= S^2 \left[\sum_{i=0}^{k} {x_i}^2 R\left(\frac{n_i}{N}\right)\left(\frac{N-n_i}{N}\right) - 2\sum_{i=0}^{k-1} \sum_{j=i+1}^{k} x_i x_j R\left(\frac{n_i}{N}\right)\left(\frac{n_j}{N}\right) \right]$$

$$+ R^2 \left[\sum_{i=0}^{k} {x_i}^2 S\left(\frac{n_i}{N}\right)\left(\frac{N-n_i}{N}\right) - 2\sum_{i=0}^{k-1} \sum_{j=i+1}^{k} x_i x_j S\left(\frac{n_i}{N}\right)\left(\frac{n_j}{N}\right) \right]$$

$$= \frac{S}{N^2} \left[\sum_{i=0}^{k} {x_i}^2 SRn_i(N-n_i) - 2\sum_{i=0}^{k-1} \sum_{j=i+1}^{k} x_i x_j SRn_i n_j \right]$$

$$+ \frac{R}{N^2} \left[\sum_{i=0}^{k} {x_i}^2 SRn_i(N-n_i) - 2\sum_{i=0}^{k-1} \sum_{j=i+1}^{k} x_i x_j SRn_i n_j \right]$$

$$= \frac{SR}{N} \left[\sum_{i=0}^{k} {x_i}^2 \, n_i(N-n_i) - 2\sum_{i=0}^{k-1} \sum_{j=i+1}^{k} x_i x_j n_i n_j \right].$$

For a sufficiently large N, we then have:

$$\frac{U}{SD(U)} = \frac{\sum_{i=0}^{k} x_i \,(Sr_i - Rs_i)}{\sqrt{\frac{N(N-R)}{N}\left[\sum_{i=0}^{k} x_i{}^2\, n_i(N-n_i) - 2\sum_{i=0}^{k-1}\sum_{j=i+1}^{k} x_i x_j n_i n_j\right]}} \sim N(0,1).$$

According to the results by Sasieni's (1997), we have

$$X_G^2 = \frac{U^2}{var(U)}$$

$$= \frac{\left[\sum_{i=0}^{k} x_i \,(Sr_i - Rs_i)\right]^2}{\frac{N(N-R)}{N}\left[\sum_{i=0}^{k} x_i{}^2\, n_i(N-n_i) - 2\sum_{i=0}^{k-1}\sum_{j=i+1}^{k} x_i x_j n_i n_j\right]} \sim X_1^2$$

We have two genotype tables (male, female) known to independent and identically distributed and each follows normal distribution N(0, 1) approximately. Therefore, we combined these two statistics to make single statistic which follows normal distribution N(0, 1).

$$\frac{U_{male} + U_{female}}{\sqrt{var(U_{male}) + var\left(U_{female}\right)}} \sim N(0,1)\ \ under\ H_0.$$

Combining P-value by Brown's method

There are several statistics depending on different coding strategies which explain various X-chromosome inactivation process in female and several p-values corresponding to each statistics. To calculate combining p-value from several p-values, we should calculate the correlation between those statistics. Assume that X_s denotes sth set of scores and i equals to 1 to 2 for male, 1 to k for female (k = number of d (skewedness)). Then the variance-covariance matrix Φ for each statistics can be expressed like that:

$$\Phi_{male} = \begin{pmatrix} \mathrm{var}(X_1) & \mathrm{cov}(X_1, X_2) \\ \mathrm{cov}(X_2, X_1) & \mathrm{var}(X_2) \end{pmatrix}, \ \Phi_{female} = \begin{pmatrix} \mathrm{var}(X_1) & \mathrm{cov}(X_1, X_2) & \cdots & \mathrm{cov}(X_1, X_k) \\ \mathrm{cov}(X_2, X_1) & \mathrm{var}(X_2) & \cdots & \mathrm{cov}(X_2, X_k) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{cov}(X_k, X_1) & \mathrm{cov}(X_k, X_2) & \cdots & \mathrm{var}(X_k) \end{pmatrix}$$

For female,

$$\mathrm{var}(X_s) = 2p(1-p)\{2(d-1)^2 p^2 + 2(1-d^2)p + d^2\}$$
$$\mathrm{cov}(X_s, X_{s^*}) = 2p\{d_s d_{s^*}(1-p) + 2p - 2d_s d_{s^*} p(1-p)^2 - 2(d_s + d_{s^*})p^2(1-p) + 2p^3\}$$
$$\text{where } d_s, d_{s^*} \in [0,2],\ d_s \neq d_{s^*}$$

For male,

$$\mathrm{var}(X_i) = 4p(1-p)\ \left(XCI\right),\ p(1-p)\ (XCI - E)$$
$$\mathrm{cov}(X_s, X_{s^*}) = 2p^2(1-p^2),\ \text{where}\ X_s \in \{0,\ 1\}, X_{s^*} \in \{0,\ 2\}$$

By using correlation of statistics, we can obtain the combining p-value when statistics are correlated [Morton B. Brown, 1975]. Assume that p_s is sth p-value and under the null hypothesis summation of $-2log_e p_s$ follows a chi-square variate with 2 degrees of freedom.

Define

$$X^2 = \sum_{s=1}^{k} -2\log_e p_s .$$

Under the null hypothesis, the mean of X^2 is 2k and the variance can be obtained by

$$\sigma^2(X^2) = \sum_{s}\sum_{s^*} \mathrm{cov}(-2\log_e p_s, -2\log_e p_{s^*})$$
$$= \sum_{s} \mathrm{var}(-2\log_e p_s) + 2\sum_{s<s^*}\sum \mathrm{cov}(-2\log_e p_s, -2\log_e p_{s^*})$$
$$= 4k + 2\sum_{s<s^*}\sum \mathrm{cov}(-2\log_e p_s, -2\log_e p_{s^*})$$

p_s and p_{s^*} are not independent, and X^2 does not follow the chi-square distribution. Distribution of X^2 can be approximated by Morton B. Brown's approach [Morton B. Brown, 1975]. If we let

$$E(X^2) = cf \text{ and } \sigma^2(X^2) = 2c^2 f.$$

X^2 can be approximated by $c \cdot \chi^2(\mathrm{df} = f)$ [Morton B. Brown, 1975]. f and c can be obtained by

$$f = 2[E(X^2)]^2 / \sigma^2(X^2) \text{ and } c = \sigma^2(X^2) / \{2E(X^2)\}.$$

Minimum P-value

P-values can be calculated for different biological processes, and the minimum p-value can be used as a test statistic. The asymptotic distribution of minimum P-value can be calculated by considering correlations among statistics. Suppose that T_{sk} $(s = 1, \dots, n,\ k = 1, \dots, m)$ means a statistics which is sth scenario with kth coding strategy and t_c $(c = 1, \dots, sk)$ denotes the observed statistics of T_{sk}. Then if we let t_{max} = max(t_1, … , t_{sk}), P-value for minimum P-value test statistic can be obtained by

$$\begin{aligned} \mathrm{P}_{\min} &= \mathrm{P}\{\max(\,|T_{s1}|, |T_{s2}|, \cdots, |T_{sk}|) > t_{\max}\} \\ &= 1 - \mathrm{P}\{\max(\,|T_{j1}|, |T_{j2}|, \cdots, |T_{jk}|) < t_{\max}\} \\ &= 1 - \mathrm{P}\{\ |T_{j1}| < t_{\max}, |T_{j2}| < t_{\max}, \cdots, |T_{jk}| < t_{\max})\ \}. \end{aligned}$$

Simulation Studies

The simulation model

In our simulation studies, we considered 1,000 males and 1,000 females with

1,000 cases and 1,000 controls. We assumed that the disease prevalence and disease allele frequency were 0.2 and the disease status for each individuals was generated with the liability threshold model. Underlying liability score of individual i with genotype X_i were defined by summing the main genetic effect, $X_i\beta$, polygenic additive effect, P_i, and random error, ε_i, as follows :

$$y_i = \beta_0 + X_i\beta + P_i + \varepsilon_i,\ P \sim \mathrm{N}(0, {\sigma_p}^2),\ \varepsilon \sim \mathrm{N}(0, \sigma^2),\ \sigma^2 = 1$$

β_0 was assumed to be 0, and σ^2 were assumed to be 1.. For the polygenic effect and random errors were denoted by σ_p^2 and σ^2, respectively, and σ^2 were assumed to be 1. We assumed that the heritability, h^2, and, for empirical power calculation, the relative proportion of variance attributable to the disease genotype, ${h_a}^2$, are 0.5, and 0.005, respectively. Then we obtain $\beta = 0.1767$ and $\sigma_P = 1$ from

$$h^2(\text{heritability}) = \frac{2P(1-P)\beta^2 + \sigma_P^2}{2P(1-P)\beta^2 + \sigma_P^2 + \sigma^2} = 0.5$$

$$\text{and,} \qquad h_a^2 = \frac{2P(1-P)\beta^2}{2P(1-P)\beta^2 + \sigma_P^2 + \sigma^2} = 0.005.$$

Under the null hypothesis, ${h_a}^2$ was set to 0, and β became 0. Once the underlying liabilities of subjects were generated, they were transformed to disease statuses; subjects became affected if their liability scores were larger than the threshold, and otherwise, they were considered as unaffected.

We are considering six scenarios according to several X-chromosome inactivation process in female: five scenarios for random XCI and skewed XCI and the other one for escaped XCI. First, we used genotype coding X={0, 0.2 (or 0.5), 2} for female to denote genotypes aa, Aa, AA, respectively, a scenario where 10%

(25%) of the cells have the risk allele and the other 90% (75%) of the cells have the normal allele. We also considered X={0, 1, 2} for female, reflecting 50% of the cells having the risk allele and the other 50% of the cells having the normal allele. And then we considered X={0, 1.5 (or 1.8), 2} for female, which means 75% (90%) of the cells have the risk allele and 25% (10%)of the cells have the normal allele. The X-chromosome coding for male about three situations is X={0, 2} to maintain balancedness with female's gene expression level. Final scenario is escaped XCI in female where coding for female X-chromosome is X={0, 1, 2} and X={0, 1} for male. Therefore, we generated six data sets with various scenarios and applied six coding method to each data sets. Finally, we make a combining p-value and minimum p-value by using these six statistics to make more robust statistics.

Results

To assess the robustness of the proposed method against the various biological model, we examined the performance of proposed methods under six different biological models including random XCI, skewed XCI, and escape from XCI. First we evaluate the statistical validity with empirical type-1 error estimates. The Figure 1 and Table 1 show that the type 1 errors of proposed methods are well preserved under the null hypothesis at nominal significance level of 0.05. Our simulation results show that the proposed methods are always the most efficient. There may be a considerable loss of power when the data does not correspond with the true underlying coding method (Table 5). Considering the $X_{d=1}$ case, the powers are similar to the case of

Table 3. Empirical type1 error for various scenarios with several coding method,

combined p-value (P_{com}) and minimum p-value (P_{min}) (α =0.05, based on 5,000 replications)

Type1_error	X_E	$X_{d=1}$	P_{com}	P_{min}
XCI_E	0.0506	0.0512	0.0506	0.0512
$XCI_{0.2}$	0.0514	0.055	0.0513	0.0542
$XCI_{0.5}$	0.0556	0.0522	0.0556	0.0525
XCI_1	0.0524	0.0526	0.057	0.0548
$XCI_{1.5}$	0.0511	0.0504	0.0532	0.0522
$XCI_{1.8}$	0.0509	0.0532	0.0512	0.0513

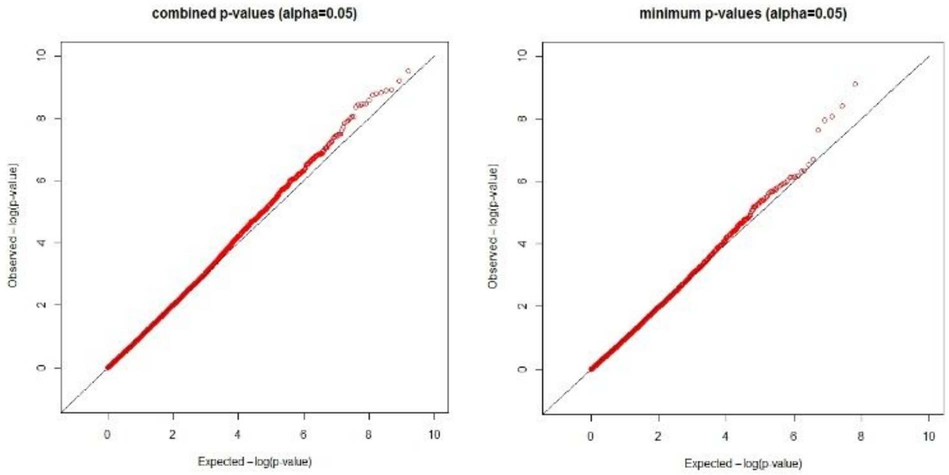


Figure 1. Q-Q plot of P-values from Cochran-Armitage statistics, combined P-values and minimum P-values, plotted on $-log_{10}$ scale at significance level 0.05.

combining p-value and minimum p-value in the most generic model. However, in case of the escaped XCI our proposed methods are evidently robust (Figure 2 and Table 5). Furthermore, since we don't know exactly about the mode of inheritance

Table 4. Coding method for random variable X for various scenarios (coding= {male; female})

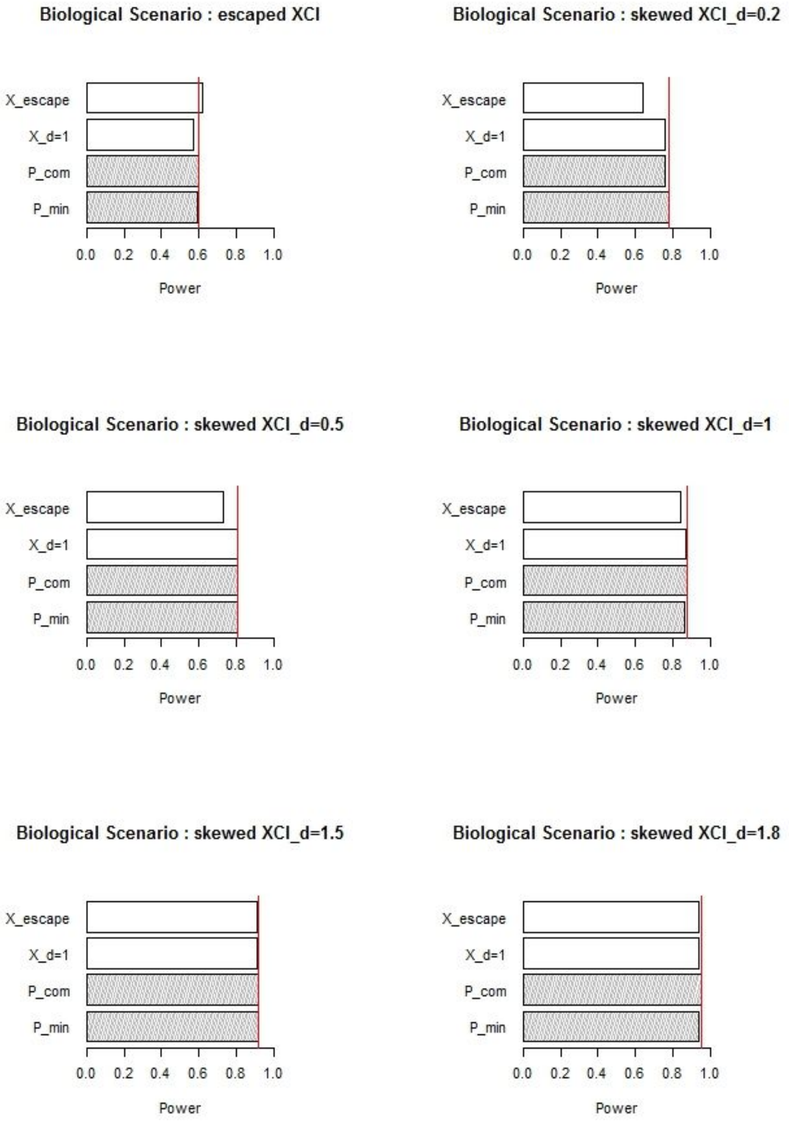
Model	**Scenarios**	**Coding**
XCI_E	XCI was not observed in female	$X_E = \{0,1;0,1,2\}$
$XCI_{0.2}$	XCI occurred with d=0.2 in female	$X_{d=0.2} = \{0,2;0,0.2,2\}$
$XCI_{0.5}$	XCI occurred with d=0.5 in female	$X_{d=0.5} = \{0,2;0,0.5,2\}$
XCI_1	XCI occurred with d=1 in female	$X_{d=1} = \{0,2;0,1,2\}$
$XCI_{1.5}$	XCI occurred with d=1.5 in female	$X_{d=1.5} = \{0,2;0,1.5,2\}$
$XCI_{1.8}$	XCI occurred with d=1.8 in female	$X_{d=1.8} = \{0,2;0,1.8,2\}$

Table 5. Empirical power for various scenarios with several coding method, combined p-value (P_{com}) and minimum p-value (P_{min}) (α =0.05, based on 5,000 replications)

	X_E	$X_{d=1}$	P_{com}	P_{min}
XCI_E	0.7798	0.7358	0.7576	0.7486
$XCI_{0.2}$	0.8234	0.905	0.9116	0.9116
$XCI_{0.5}$	0.8688	0.9292	0.9328	0.9256
XCI_1	0.9438	0.9628	0.9622	0.9548
$XCI_{1.5}$	0.9804	0.9814	0.9846	0.9796
$XCI_{1.8}$	0.991	0.99	0.9926	0.9914

and the choice of a set of scores is not clear, it would be good choice for case-control association study with X-chromosome to use our proposed approach. Averagely, our proposed method is more powerful than the worst coding strategy about 5% when d vary from 0 to 1 and less powerful than the true corresponding coding strategy about 0.7% at significance level of 0.05. there is no big difference when the d greater than 1.

Figure 2. Empirical power of various generic model with several coding method



and the proposed approach ($\alpha = 0.05$) (skewed XCI_d=1 same as random XCI)

Discussion

The biological process for X-linked variants is complicated and it may partially explain the relatively less finding of X-linked disease susceptibility loci from GWAS. Since we can't be sure about the XCI process in female, a set of score is difficult to be chosen. In this report, we proposed a new association test which can account for various plausible biological models. Our simulation studies indicate that the proposed approach is not always more robust than the other coding strategy but is always the second best for various biological models. Even if there was loss of statistical power in the proposed method, it was little and more reasonable than the other coding methods. This is because that we don't know real biological process for X-chromosome markers in our body. Therefore, we can conclude that the proposed approach is robust against the various XCI processes for testing the association of X-linked SNPs with the disease of interest. We expect that the new approach makes it useful for GWAS with X-chromosome. Despite these advances, the limitation of association approach we proposed for the analysis of X-linked markers is only appropriate for independent samples. We can't apply this method to family data which consisted of related individuals. Consequently, we will develop this method or case-control association analysis of X-linked variants when the individuals are related like family data. To this end, we should consider kinship coefficient matrix for an X-chromosome marker for family relationship which is difference between X-chromosome and autosomal markers. Lange [Lange et al., 1976] calculated X-chromosome kinship coefficients for related individuals. There are some papers about analysis for X-linked markers with related samples.

However, the approach [Thormthon et al., 2012] previously has been proposed just take account of random XCI and escaped from XCI exclusive of skewed XCI. Therefore, our next assignment to solve is making a new statistics which can consider skewed XCI besides random XCI and escaped from XCI with related individuals and we expect that it would be lead to better robust statistics.

Reference

Amos-Landgraf JM, Cottle A, Plenge RM, Friez M, Schwartz CE, Longshore J, Willard HF. 2006. X chromosome-inactivation patterns of 1,005 phenotypically unaffected females. *Am J of Hum Genet. 2006;79:493–499.*

Belmont JW. 1996. Genetic control of X inactivation and processes leading to X-inactivation skewing. *Am J of Hum Genet. 1996;58:1101–1108.*

Busque L, Paquette Y, Provost S, Roy DC, Levine RL, Mollica L, Gilliland DG. 2009. Skewing of X-inactivation ratios in blood cells of aging women is confirmed by independent methodologies. *Blood. 2009;113:3472–3474.*

Chagnon P, Provost S, Belisle C, Bolduc V, Gingras M, Busque L. 2005. Age-associated skewing of X-inactivation ratios of blood cells in normal females: a candidate-gene analysis approach. *Exp Hematol. 2005;33:1209–1214.*

Freidlin B , Zheng G, Li Z, Gastwirth JL. 2002. Trend Tests for Case-Control Studies of Genetic Markers: Power, Sample Size and Robustness. *Hum Hered. 2002;53(3):146-52.*

Jian Wang, Robert Yu and Sanjay Shete. 2014. X-chromosome Genetic Association Test Accounting for X-inactivation, Skewed X-Inactivation, and Escape from X-Inactivaton. *Version of Record online: 8 JUL 2014 DOL: 10.1002/gepi.21814*

Morton B. Brown, 1975. A Method for Combining Non-Independent, One-Sided Tests of Significance. *Biometrics. Vol. 31, No. 4, pp 987-992*

Minks J, Robinson WP, Brown CJ. 2008. A skewed view of X chromosome inactivation. *J Clin*

Invest. 2008;118:20–23.

Plenge RM, Stevenson RA, Lubs HA, Schwartz CE, Willard HF. 2002. Skewed X-chromosome inactivation is a common feature of X-linked mental retardation disorders. *Am J Hum Genet. 2002;71:168–173.*

Struewing JP, Pineda MA, Sherman ME, Lissowska J, Brinton LA, Peplonska B, Bardin

Mikolajczak A, Garcia-Closas M. 2006. Skewed X chromosome inactivation and early-onset breast cancer. *J Med Genet. 2006;43:48–53.*

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. 2014 Sep;38(6):483-93. doi: 10.1002/gepi.21814. Epub 2014 Jul 8.*

Wellek S, Ziegler A. 2012. Cochran-Armitage Test versus Logistic Regression in the Analysis of Genetic Association Studies. *Hum Hered. 73(1):14-7. doi: 10.1159/000334085. Epub 2011 Dec 30.*

Willard HF. 2000. The sex chromosomes and X chromosome inactivation. In: Scriver CR, Beaudet AL, Sly WS, Valle D, Childs B, Vogelstein B, editors. The Metabolic and Molecular Bases of Inherited Disease. *New York: McGraw-Hill; 2000. pp. 1191–1221.*

Wong CC, Caspi A, Williams B, Houts R, Craig IW, Mill J. 2011. A longitudinal twin study of

skewed X chromosome-inactivation. *PLoS One. 2011;6:e17873.*

Abstract in Korean

Y염색체보다 X염색체는 세포의 발달과 생존에 필수적인 유전자를 1,000개 이상 더 함유하고 있다. 또한, 여성의 경우 X염색체에 관련된 유전자 발현량이 남성의 2배가 될 것으로 기대된다. X염색체의 유전자에 의하여 많은 단백질이 생기는 것을 막기 위해서 여성의 세포에서는 두 개 중 한 개의 X염색체가 불활성화 되는데 이를 'X염색체 불활성화' 라고 한다. 그러나 최근의 연구들은 두 개의 X염색체 중에서 하나가 불활성화 되는 것은 랜덤하게 일어나지 않으며, 불활성화가 일어나지 않는 경우도 있다는 것을 입증하였다. 이렇게 복잡한 생물학적인 현상은 효율적인 연관성분석을 어렵게하며 X염색체와 관련된 변이의 발견도 쉽지 않게 한다. 이에 따라, 본 연구에서는 X염색체의 다양한 불활성화에 대하여 combining P-value와 minimum P-value를 이용하여 효율적인 통계적 분석방법을 제시하고자 한다. 시뮬레이션을 통하여 얻은 결과는 제시된 방법이 실제 생물학적인 상태에 맞게 코딩되었을 때 보다는 조금 덜하지만 다른 코딩 방법들에 비해 효과적임을 입증하였다.

주요어 : X염색체 불활성화, X염색체 관련 유전 변이, X염색체 연관성 분석

학 번 : 2014-23378