Abstract

Introduction

Recently, three research groups simultaneously proposed a new kind of F_{st}-like approaches in detecting loci under selection in structured populations (Chen *et al.*, 2016; Duforet-Frebourg *et al.*, 2016; Kevin J. Galinsky *et al.*, 2016). Chen *et al* (2016) further found a simple-marker regression form of the method, and because of its similarity to genome-wide association studies (GWAS), we proposed the name EigenGWAS. The EigenGWAS has also been adopted statistical analysis packages, such as pcadpt (Luu, 2017) and sommer (Covarrubias-Pazaran, 2016).

Compared with the conventional F_{st}, which asks a cutting-clear definition for subpopulation but probably unavailable especially for continuous sampling for the population of question, the newly EigenGWAS can bypass this cumbersome task. These methods found their applications in detecting loci under selection, such as in human genetics (Kevin J. Galinsky *et al.*, 2016; Chaput *et al.*, 2014; Shen *et al.*, 2016; Liu *et al.*, 2018; Lee *et al.*, 2017), ecology (Kim *et al.*, 2017; Bosse *et al.*, 2017; Armstrong *et al.*, 2018), breeding population (Ma *et al.*, 2016; Liu *et al.*, 2017; Zhao *et al.*, 2018), and other applications – especially various re-sequencing populations that are embarking.

We have observed and analyzed more data and gained further understanding of the method, in this study we are going to help further refine the application of the new method as below:

i) Reconcile the three papers aforementioned (Chen *et al.*, 2016; Duforet-Frebourg *et al.*, 2016; Kevin J Galinsky *et al.*, 2016), in which EigenGWAS will be a reasonably elementary form of the three methods.

ii) For the one-degree-of-freedom chi-square test that is included in testing the statistical significance of the loci under selection, non-centrality parameter of the test statistic is defined. Consequently, statistical power of the method can be conducted.

iii) Technical adjustment for population structure is verified by choosing between eigenvalue, which will be revealed having a mixture distribution and consequently not always suitable for the correction of population structure, and genomic inflation factor (Devlin and Roeder, 1999).

iv) Caveats for incompletely sampled populations, especially for the population under strong selection, is discussed, and empirical evidence to characterize those populations is also presented.

For the brevity of the text, we will use EigenGWAS to represent the F_{st}-like approaches.

Materials and methods

Characterizing the sample

Given the genotypic matrix $\boldsymbol{X}$, an $n \times m$ genotypic matrix in which n is the sample size and m the number of markers. By setting the current population as the reference population, we have the additive genomic

relation matrix $\boldsymbol{G} = \frac{1}{m}\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{X}}^{\boldsymbol{T}}$, in which $\widetilde{\boldsymbol{X}}$ represents the standardized form of $\boldsymbol{X}$; in particular between individual i and j, $\boldsymbol{G}_{ij} = \frac{1}{m}\sum_{l=1}^{m}\frac{(x_{il}-2\hat{p}_l)(x_{jl}-2\hat{p}_l)}{2\hat{p}_l(1-\hat{p}_l)}$. This matrix has been widely used in genetic analysis, such as genetic prediction (VanRaden, 2008).

Two useful statistics can be inferred from $\boldsymbol{G_o}$, the off-diagonal elements of the $\boldsymbol{G}$ matrix. In theory, if it is a randomly sampled population, $E(\boldsymbol{G}_o) = -\frac{1}{n}$, and upon how inbred the samples are, the lower bound of $E(\boldsymbol{G}_o)$ is $-\frac{2}{n}$. So, we can have an estimate of the samples, $n_e = -\frac{1}{E(\boldsymbol{G}_o)}$.

In addition, the variation of $var(\boldsymbol{G}_o) = \frac{m+\sum_{l_1=1}^{m}\sum_{l_2\neq l_1}^{m}\rho_{l_1l_2}^2}{m^2}$, in which $\rho_{l_1l_2}^2$ is the squared Pearson's correlation for a pair of markers—the most commonly used metric for the linkage disequilibrium (LD) of a pair of markers (Devlin and Risch, 1995). We can further define $m_e = \frac{1}{var(\boldsymbol{G}_o)}$ (Chen, 2014). It is obviously that m_e is between 1, every marker is in full LD between each other, and m, no LD between any pair of markers. It has been well-known in evolution that the when there is hitch-hicking effect the LD will be increased for the markers within the genomic region under selection (Maynard and Haigh, 1974).

So, the mean and variance of $\boldsymbol{G}_o$ can help elucidate some basic characters of the sample of study. It is expected for inbred populations, such as Arabidopsis (Alonso-Blanco *et al.*, 2016), $n_e \approx \frac{n}{2}$, while for random mating population, $n_e \approx n$.

The non-centrality parameter for EigenGWAS

After eigenvalue decomposition analysis of $\boldsymbol{G}$, we have two matrices $\boldsymbol{E}$, which is an $m \times m$ matrix, and $\boldsymbol{\Lambda}$ the diagonal matrix for the ordered eigenvalues. A linear model, EigenGWAS, can be constructed as

$\boldsymbol{E}_k = \mu + \beta_l\boldsymbol{x}_l + e$ (Eq 1)

in which $\boldsymbol{E}_k$ is the k^{th} eigenvector/column, μ the grand mean, $\boldsymbol{x}_l$ the l^{th} SNP/column, and e the residual. We are interested in the regression coefficient β_l. In Appendix, we can derive the genetic interpretation of β_l.

Under the null distribution of no association, $\frac{\widehat{\beta}_l^2}{\sigma_{\beta_l}^2} \sim \chi_1^2$. When the locus is associated with $\boldsymbol{E}_k$, the NCP is approximately $4n\omega_1\omega_2\frac{(p_{l|1}-p_{l|2})^2}{2p_l(1-p_l)}$. $\omega_1 = \frac{n_1}{n}$ and $\omega_2 = \frac{n_2}{n}$, n_1 and n_2 are the numbers of samples that have positive and negative values in $\boldsymbol{E}_k$.

There are many kinds of F_{st}, we only introduce two forms here that is closed related to EigenGWAS. $F_{st}^{N} = \frac{(p_{l|1}-p_{l|2})^2}{2p_l(1-p_l)}$ and $F_{st}^{N} = 2\frac{\omega_1(p_{l|1}-p_l)^2+\omega_2(p_{l|2}-p_l)^2}{p_l(1-p_l)}$. The connection between the NCP and F_{st} can be, after some algebra, established

$$NCP = \begin{cases} 4n\omega_1\omega_2 F_{st}^{N} \\ nF_{st}^{W} \end{cases}$$

Given expression above, we can learn that

- EigenGWAS only see "two subgroups" that at the two sides of the point zero.
- the power is proportional to sample size n, and maximized when $\omega_1 = \omega_2 = \frac{n}{2}$.
- NCP is zero if $p_1 = p_2$.

If we are interested in construct a statistical hypotheses test for whether a locus is under selection, it is better to control the background information, such as genetic drift. Because the conventional F_{st} statistic contains both genetic drift and selection, we are aim to control genetic drift.

Eigenvalue and genomic inflation factor

As the distribution of the test statistic is known, it is possible to coin the test statistic further, for example to test the selection by controlling genetic drift. For a locus of interest, F_{st} is an overall measurement for genetic differentiation, and in order to control for genetic drift, the adjustment such as genomic inflation factors can be used.

Originally, the three papers all used the k^{th} eigenvalue that is associated with $\boldsymbol{E}_k$ to adjust the test statistic, so the de facto test statistic is (Chen *et al.*, 2016; Duforet-Frebourg *et al.*, 2016; Kevin J Galinsky *et al.*, 2016)

$$n\frac{F_{st}^{W}}{\lambda_k}$$

Of note, in Duforet-Frebourg et al, it was a scan statistic and its distribution was not that well characterized. Later on, Luu et al (Luu, 2017) adopted a test statistic and has been updated it to

$$n\frac{F_{st}^{W}}{\lambda_{GC.k}}$$

in which $\lambda_{GC.k}$ is the genomic inflation factor (Devlin and Roeder, 1999). In practice, some authors also adopted $\lambda_{GC.k}$ in data analysis (Bosse *et al.*, 2017).

As observed by us (Table 1 in our paper), we found the eigenvalue and the genomic inflation factor were very similar and often highly correlated along many scanned eigenvectors, and we adopted eigenvalues as a way to control for genetic drift (Chen *et al.*, 2016). It is very interesting to see these subtle difference. Because we know both eigenvalue and genomic inflation factor can reflect population structure, what is the

consequence and implication for the respective application in data analysis. In particular, whether it will influence the statistical power because of the choice of the technical adjustment for population structure.

It is known that there is connection between F_{st} and eigenvalue, in particular for human population the largest eigenvalue often reflects genetic drift for the population geographically distributed (Patterson *et al.*, 2006; McVean, 2009). We shows that an eigenvalue is a mixture distribution, and at least can be written as two components

$$\lambda_1 \propto \eta \tilde{F}_{st} + (1-\eta)\bar{F}_{st}$$

in which η is the proportion of the loci under selection the genetic differentiation of that is $\tilde{F}_{st}$ and the rest of the loci under genetic differentiation $\bar{F}_{st}$. $\tilde{F}_{st} \geq \bar{F}_{st}$.

It could be anticipated that when η is a small number, often small for such as human population, the median of the observed χ_1^2 reflects $\bar{F}_{st}$ because the signals soak of selection signals are ranked at the high end of the observed χ_1^2. So, adjust the test statistic with λ_{GC} will help control genetic drift. However, if η is small and the selection is moderate, we are not expected to see λ_1 differs from λ_{GC} much.

Upon the genetic architecture of the population, these two adjustment of the test statistic differ dramatically. The interpretation of $\bar{F}_{st}$, upon the populations, may find its interpretation of not. For human population, the top eigenvalue often indicate geographic isolation of the samples. A very great example is great tit samples. For great tit samples, the first eigenvalue reflects the genetic isolation of the samples by the ocean, and the second eigenvalue the separation between two Netherland locations (Bosse *et al.*, 2017). Need another scan for PC2.

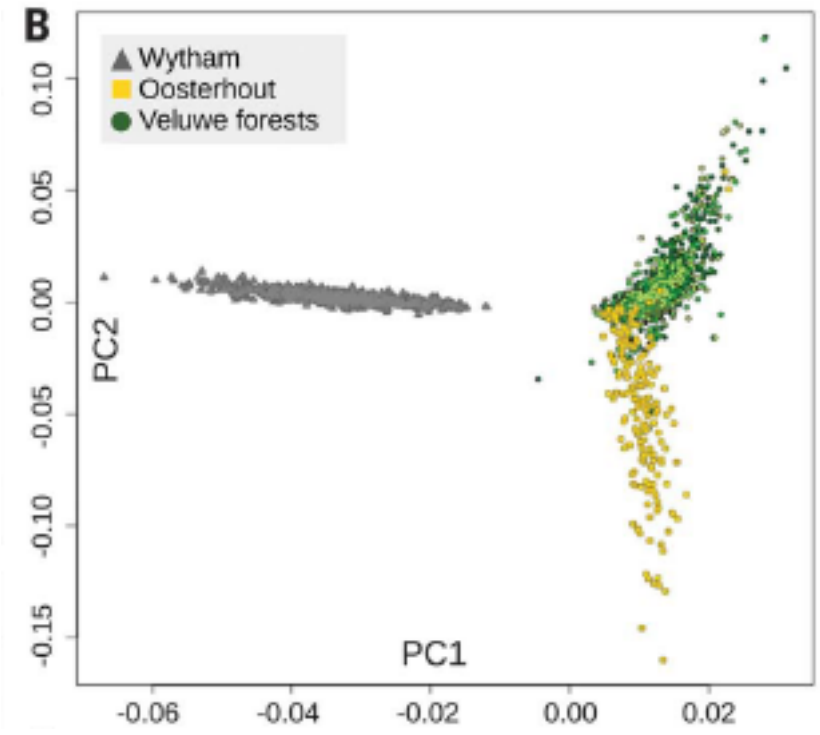


Table 1

Species	Population	n	n_e	m [maf>0.1]	m_e	$\bar{F}_{st}$	Source
Human	HapMap						
	POPRES	2,466	2,450	615,494	119,407		
	UK Biobank						
Great tit	Great tit	3015	5740.75	484370	6212.54		
Dog		1792	1447.35	164,164	47.54		
Plant	Maize HapMap3	879	902	335848	104.12		
	Arabidopsis 294	294	293	156,744	394.73		
	Arabidopsis 1135						
	3K rice	357 [which 357]	348	2,702,622	8		

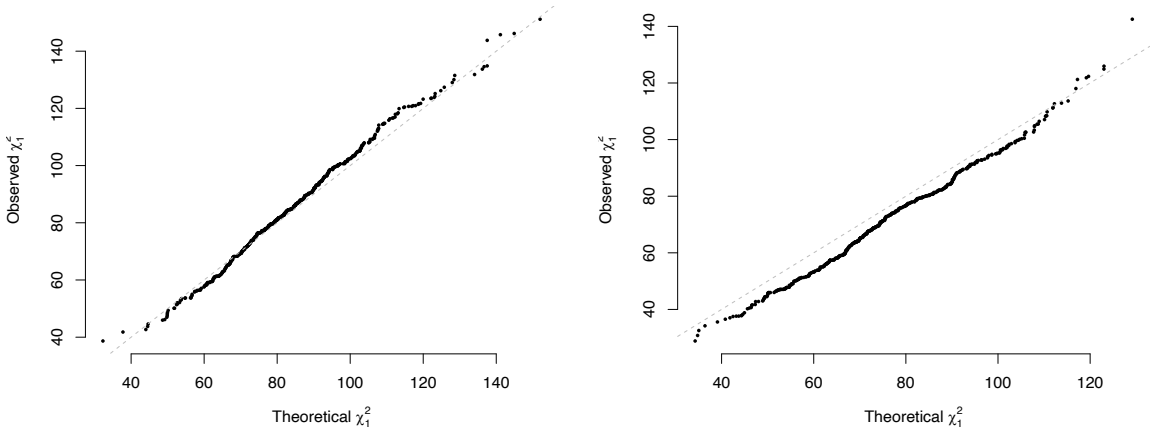
For eigenvalues $\lambda_1 > \lambda_2$, but it is not necessary the case for the genomic inflation factors for example maize HapMap3.

Simulation

Validation of the NCP

We validated our NCP by the simulation below. We simulated a locus with allele frequency of 0.4 in subpopulation 1 and 0.6 in subpopulation 2, and the subpopulations had equal sample size of 500. According to the NCP was 82.51, and we replicated the sample 500 times. Then in a quantile-quantile plot, we compared the observed chisq test statistic with the one sampled from a chisq distribution with NCP of 82.51.

Alternatively, we changed the sample sizes to 300 vs 700, and the observed fitted the expectation well too.

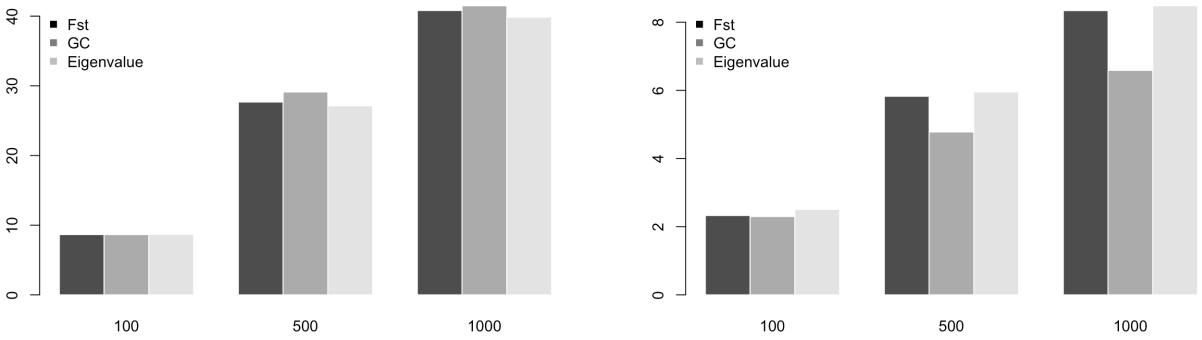


The mixture distribution of eigenvalue

We show that the distribution of the eigenvalue can be a mixture distribution. We simulated population 1 the eigenvalue was dominated by a single F_{st}, which mimicked genetic drift. The allele frequencies of the two subpopulations were sampled from $Beta(p\frac{1-F_{st}}{F_{st}}, (1-p)\frac{1-F_{st}}{F_{st}})$, in which f is the allele frequency of their

common ancestor, and $F_{st} = \frac{p_1 - p_2}{2p(1-p)}$; in the second simulation, the loci were partitioned into two sets, the first set was driven by genetic drift sampled from $Beta(p\frac{1-\bar{F}_{st}}{\bar{F}_{st}}, (1-p)\frac{1-\bar{F}_{st}}{\bar{F}_{st}})$, and the second set from $Beta(p\frac{1-\tilde{F}_{st}}{\tilde{F}_{st}}, (1-p)\frac{1-\tilde{F}_{st}}{\tilde{F}_{st}})$.

For the simulated samples, we conducted EigenGWAS. Their F_{st}, λ_1, and $\lambda_{GC.1}$ as illustrated below



Real data analysis

UKBiobank

Discussion

Acknowledgements

This study was supported by the National Natural Science Foundation of China (31771392 to G-B C, and 31871707 to H-M X) and Zhejiang Provincial People's Hospital Research Foundation (ZRY2018A004 to G-B C). The funders played no role in experimental data and data analysis.

Reference

Alonso-Blanco,C. *et al.* (2016) 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. *Cell*, **166**, 481–491.

Armstrong,C. *et al.* (2018) Genomic associations with bill length and disease reveal drift and selection across island bird populations. *Evol. Lett.*, **2**, 22–36.

Bosse,M. *et al.* (2017) Recent natural selection causes adaptive evolution of an avian polygenic trait. *Science*, **358**, 365–368.

Chaput,J.-P. *et al.* (2014) Findings from the Quebec Family Study on the Etiology of Obesity: Genetics and Environmental Highlights. *Curr. Obes. Rep.*, **3**, 54–66.

Chen,G.-B. *et al.* (2016) EigenGWAS: finding loci under selection through genome-wide association studies of eigenvectors in structured populations. *Heredity*, **117**, 51–61.

Chen,G.-B. (2014) Estimating heritability of complex traits from genome-wide association studies using IBS-based Haseman-Elston regression. *Front. Genet.*, **5**, 107.

Covarrubias-Pazaran,G. (2016) Genome-Assisted prediction of quantitative traits using the R package sommer. *PLoS ONE*, **11**, e0156744.

Devlin,B. and Risch,N. (1995) A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics*, **29**, 311–22.

Devlin,B. and Roeder,K. (1999) Genomic control for association studies. *Biometrics*, **55**, 997–1004.

Duforet-Frebourg,N. *et al.* (2016) Detecting genomic signatures of natural selection with principal component analysis: Application to the 1000 genomes data. *Mol. Biol. Evol.*, **33**, 1082–1093.

Galinsky,K.J. *et al.* (2016) Fast principal components analysis reveals independent evolution of ADH1B gene in Europe and East Asia. *Am. J. Hum. Genet.*, **98**, 456–472.

Galinsky,K.J. *et al.* (2016) Population Structure of UK Biobank and Ancient Eurasians Reveals Adaptation at Genes Influencing Blood Pressure. *Am. J. Hum. Genet.*, **99**, 1130–1139.

Kim,K.-W. *et al.* (2017) A sex-linked supergene controls sperm morphology and swimming speed in a songbird. *Nat. Ecol. Evol.*

Lee,S.H. *et al.* (2017) Using information of relatives in genomic prediction to apply effective stratified medicine. *Sci. Rep.*, **7**, 42091.

Liu,S. *et al.* (2018) Genomic Analyses from Non-invasive Prenatal Testing Reveal Genetic Associations, Patterns of Viral Infections, and Chinese Population History. *Cell*, **175**, 347–359.

Liu,Z. *et al.* (2017) Comparison of Genetic Diversity between Chinese and American Soybean (Glycine max (L.)) Accessions Revealed by High-Density SNPs. *Front. Plant Sci.*, **8**, 2014.

Luu,K. (2017) pcadapt : an R package to perform genome scans for selection based on principal component analysis. *Mol. Ecol. Resour.*, **17**, 67–77.

Ma,X. *et al.* (2016) Genome-Wide Association Study for Plant Height and Grain Yield in Rice under Contrasting Moisture Regimes. *Front. Plant Sci.*, **7**, 1801.

Maynard,J. and Haigh,J. (1974) The hitch-hiking effect of a favourable gene. *Genet. Res.*, **23**, 23–35.

McVean,G. (2009) A genealogical interpretation of principal components analysis. *PLoS Genet.*, **5**, e1000686.

Patterson,N. *et al.* (2006) Population structure and eigenanalysis. *PLoS Genet.*, **2**, e190.

Shen,Q.K. *et al.* (2016) Was ADH1B under Selection in European Populations? *Am. J. Hum. Genet.*, **99**, 1217–1219.

VanRaden,P.M. (2008) Efficient methods to compute genomic predictions. *J. Dairy Sci.*, **91**, 4414–4423.

Weir,B.S. and Cockerham,C.C. (1984) Estimating F-Statistics for the Analysis of Population Structure. *Evolution*, **38**, 1358–1370.

Zhao,Q. *et al.* (2018) Identifying Genetic Differences between Dongxiang Blue-shelled and White Leghorn Chickens using Sequencing Data. *G3*, **8**, 469–476.

NCP for the additive model

For random mating populations

For linear model $\widetilde{\boldsymbol{E}}_k = \mu + \beta_l \boldsymbol{x}_l + e$,

In which $\widetilde{\boldsymbol{E}}_k$ the standardized eigenvector of interest, $\boldsymbol{x}_l$ codes for the dominance effect, β_l the regression coefficien, and e the residual of the model. The regression coefficient can be expressed as

$$E(\beta_l) = \frac{cov(\boldsymbol{x}_l, \widetilde{\boldsymbol{E}}_k)}{var(\boldsymbol{x}_l)}$$

		x_l			Marginal probability	
	$\widetilde{\boldsymbol{E}}_k$	aa	Aa	AA		
Subpop 1	$\frac{\sqrt{\frac{n_2}{n_1}}}{1+\omega_1\sigma_1^2+\omega_2\sigma_2^2}$	2	1	0		$\omega_1 = \frac{n_1}{n_1+n_2}$
		p_1^2	$2p_1q_1$	q_1^2	$1-\gamma_1$	
		p_2^2	$2p_2q_2$	q_2^2	γ_2	
Subpop 2	$-\frac{\sqrt{\frac{n_1}{n_2}}}{1+\omega_1\sigma_1^2+\omega_2\sigma_2^2}$	2	1	0		$\omega_2 = \frac{n_2}{n_1+n_2}$
		p_2^2	$2p_2q_2$	q_2^2	$1-\gamma_2$	
		p_1^2	$2p_1q_1$	q_1^2	γ_1	

$$cov(y,x) = E(xy) - E(x)E(y) = E(xy)$$

$$E(xy) = \frac{\sqrt{\frac{n_2}{n_1}}}{1+\omega_1\sigma_1^2+\omega_2\sigma_2^2}[\omega_1(2p_1^2+2p_1q_1)(1-\gamma_1) + \omega_2(2p_2^2+2p_2q_2)\gamma_2]$$

$$-\frac{\sqrt{\frac{n_1}{n_2}}}{1+\omega_1\sigma_1^2+\omega_2\sigma_2^2}[\omega_2(2p_2^2+2p_2q_2)(1-\gamma_2) + \omega_1(2p_1^2+2p_1q_1)\gamma_1]$$

$$= \frac{2\sqrt{\frac{n_2}{n_1}}}{1+\omega_1\sigma_1^2+\omega_2\sigma_2^2}[\omega_1 p_1(1-\gamma_1) + \omega_2 p_2\gamma_2] - \frac{2\sqrt{\frac{n_1}{n_2}}}{1+\omega_1\sigma_1^2+\omega_2\sigma_2^2}[\omega_2 p_2(1-\gamma_2) + \omega_1 p_1\gamma_1]$$

Under the condition that $\gamma_1 = 0$ and $\gamma_2 = 0$, and the sampling variance further shrinks to zero, the NCP becomes

$$2\sqrt{\frac{n_1}{N}\frac{n_2}{N}}(p_1 - p_2)$$

		x_d			Marginal probability
	$\widetilde{\boldsymbol{E}}_k$	aa	Aa	AA	
Subpop 1	$\sqrt{\frac{n_2}{n_1}}$	2	1	0	$\omega_1 = \frac{n_1}{n_1+n_2}$
		p_1^2	$2p_1q_1$	q_1^2	
Subpop 2	$-\sqrt{\frac{n_1}{n_2}}$	2	1	0	$\omega_2 = \frac{n_2}{n_1+n_2}$
		p_2^2	$2p_2q_2$	q_2^2	
		$f_1 = \omega_1 p_1^2 + \omega_2 p_2^2$	$f_2 = \omega_1 h_1 + \omega_2 h_2$	$f_3 = \omega_1 q_1^2 + \omega_2 q_2^2$	

$$cov(y,x) = E(xy) - E(x)E(y) = E(xy)$$

$$E(xy) = \omega_1\sqrt{\frac{n_2}{n_1}}(2p_1^2+2p_1q_1) - \omega_2\sqrt{\frac{n_1}{n_2}}(2p_2^2+2p_2q_2) = 2\sqrt{\frac{n_1}{N}\frac{n_2}{N}}(p_1-p_2)$$

we have

$$E(\beta_l) \approx \frac{2\sqrt{\omega_1\omega_2}(p_1 - p_2)}{2\bar{p}\bar{q}}$$

The sampling variance of b_d is $\sigma_{\beta_l} = \sqrt{\frac{\sigma_y^2 - \beta_l^2\sigma_{x_l}^2}{(N-1)\sigma_{x_l}^2}}$, so the z-score statistic is

$$z \approx \sqrt{(N-1)\omega_1\omega_2}\frac{2(p_1 - p_2)}{\sqrt{2\bar{p}\bar{q}}}$$

and $z^2 \sim \chi_1^2$, with ncp of $4(N-1)\omega_1\omega_2\frac{(p_1-p_2)^2}{2\bar{p}\bar{q}}$. In comparison, for the additive model, the ncp approximates to $4n\omega_1\omega_2\frac{(p_{l|1}-p_{l|2})^2}{2p_lq_l}$.

For inbred lines

For linear model $\widetilde{\boldsymbol{E}}_k = \mu + \beta_l\boldsymbol{x}_l + e$,

		x_l		Marginal probability
	$\widetilde{\boldsymbol{E}}_k$	aa	AA	
Subpop 1	$\sqrt{\frac{n_2}{n_1}}$	2	0	$\omega_1 = \frac{n_1}{n_1+n_2}$
		p_1	q_1	
Subpop 2	$-\sqrt{\frac{n_1}{n_2}}$	2	0	$\omega_2 = \frac{n_2}{n_1+n_2}$
		p_2	q_2	
		$f_1 = \omega_1p_1 + \omega_2p_2$	$f_2 = \omega_1q_1 + \omega_2q_2$	

$$E(xy) = -\omega_1\sqrt{\frac{n_2}{n_1}}2p_1 + \omega_2\sqrt{\frac{n_1}{n_2}}2p_2 = 2\frac{\sqrt{n_1n_2}}{N}(p_1 - p_2)$$

$$E(\beta_l) = \frac{E(xy)}{4pq} = \frac{2\frac{\sqrt{n_1n_2}}{N}(p_1 - p_2)}{4pq}$$

$$\sigma_{\beta_l} = \sqrt{\frac{\sigma_y^2 - \beta_l^2\sigma_{x_l}^2}{(n-1)\sigma_{x_l}^2}}$$

$$z \approx \sqrt{(n-1)\omega_1\omega_2}\frac{2(p_1 - p_2)}{\sqrt{4\bar{p}\bar{q}}}$$

then the NCP approximates to $(n-1)\omega_1\omega_2\frac{(p_1-p_2)^2}{\bar{p}\bar{q}}$.

So, the NCP for inbred population is the half of that of a random mating population.

NCP for dominance model

For linear model $y = \mu + \beta_d x_d + e$,

In which y the standardized eigenvector of interest, x_d codes for the dominance effect, b_d the regression coefficient for the dominance effect, and e the residual of the model. The regression coefficient can be expressed as

$$E(\beta_d) = \frac{cov(x,y)}{var(x)}$$

		x_d			Marginal probability
	y	aa	Aa	AA	
Pop 1	$-\sqrt{\frac{n_1}{n_2}}$	0	1	0	$\omega_1 = \frac{n_1}{N}$
		p_1^2	$2p_1q_1$	q_1^2	
Pop 2	$\sqrt{\frac{n_2}{n_1}}$	0	1	0	$\omega_2 = \frac{n_2}{N}$
		p_2^2	$2p_2q_2$	q_2^2	
		$f_1 = \omega_1 p_1^2 + \omega_2 p_2^2$	$f_2 = \omega_1 h_1 + \omega_2 h_2$	$f_3 = \omega_1 q_1^2 + \omega_2 q_2^2$	

$$cov(y,x) = E(xy) - E(x)E(y) = E(xy)$$

$$E(xy) = \omega_1 \sqrt{\frac{n_2}{n_1}} 2p_1q_1 - \omega_2 \sqrt{\frac{n_1}{n_2}} 2p_2q_2 = \sqrt{\frac{n_1}{N}\frac{n_2}{N}}(2p_1q_1 - 2p_2q_2)$$

We have $E(x_d) = f_2 = \omega_1$, $var(x_d) = f_2(1-f_2)$

$$E(\beta_d) = \frac{\sqrt{\omega_1\omega_2}(h_1 - h_2)}{f_2(1-f_2)}$$

in which $h_l = 2p_lq_l$.

The sampling variance of b_d is $\sigma_{\beta_d} = \sqrt{\frac{\sigma_y^2 - \beta_d^2\sigma_{x_d}^2}{(n-1)\sigma_{x_d}^2}}$, so the z-score statistic is

$$z = \frac{\frac{\sqrt{\omega_1\omega_2}(h_1 - h_2)}{f_2(1-f_2)}}{\sqrt{\frac{\sigma_y^2 - b_d^2\sigma_{x_d}^2}{(n-1)\sigma_{x_d}^2}}} \approx \sqrt{(n-1)\omega_1\omega_2}\frac{(h_1 - h_2)}{\sqrt{f_2(1-f_2)}}$$

and $z^2 \sim \chi_1^2$, with ncp of $(n-1)\omega_1\omega_2\frac{(h_1-h_2)^2}{f_2(1-f_2)}$. In comparison, for the additive model, the ncp approximates to $4n\omega_1\omega_2\frac{(p_{l|1}-p_{l|2})^2}{2p_lq_l}$.