Distribution Agreement

In presenting this thesis as a partial fulfillment of the requirements for a degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis in whole or in part in all forms of media, now or hereafter now, including display on the World Wide Web. I understand that I may select some access restrictions as part of the online submission of this thesis. I retain all ownership rights to the copyright of the thesis. I also retain the right to use in future works (such as articles or books) all or part of this thesis.

Se Jun Park March 24, 2025

Exploring X-linked Contributions to Sex-Specific Prevalence Differences in Autism Spectrum Disorder (ASD)

by

Se Jun Park

David J. Cutler
Adviser

Department of Biology

David J. Cutler
Committee Member

Michal Arbilly
Committee Member

Hojin Kim
Committee Member

2025

Exploring X-linked Contributions to Sex-Specific Prevalence Differences in Autism Spectrum Disorder (ASD)

By

Se Jun Park

David J. Cutler
Adviser

An abstract of a thesis submitted to the Faculty of
Emory College of Arts and Sciences of Emory University
in partial fulfillment of the requirements of the degree of
Bachelor of Science with Honors

Department of Biology

2025

Abstract

Exploring X-linked Contributions to Sex-Specific Prevalence Differences in Autism Spectrum Disorder (ASD)

By Se Jun Park

Autism spectrum disorder is a complex neurodevelopmental disorder that shows a marked sex difference with increased prevalence in males than in females. While much of the genetic risk for ASD is due to many risk genes, there is growing evidence that X-linked genes contribute prominently to the male-female difference in ASD prevalence. Thus, with respect to X-linked loci, this thesis employs a liability threshold model to better understand how X-linked loci may contribute to the sex-differentiated genetic risk of ASD. For males and females, we will derive and compare mean and variance components under three genetic models: additive, fully dominant, and fully recessive; using equations and plots to demonstrate the role of allele frequency, the effect size in a liability threshold model, as our final goal is to obtain the number of loci (n) that will need to observe the prevalence difference.

Determining the number of loci, in an additive model, if the loci have large effect sizes, a plausible estimate is around 1 to 10 loci, whereas thousands may be necessary when the effect sizes are small. For a fully dominant model, one to five loci would be sufficient if the effect sizes are large and a hundred or more loci are needed when the effect sizes are small. However, the fully recessive model yields a valid solution for the number of loci only when the effect parameter is negative, indicating that the risk-increasing allele acts in the opposite direction compared to the additive and dominant cases. In this case, a plausible range is approximately one to 100 loci for large effect sizes and above 1000 loci when the effect sizes are small.

These results suggest that sex differentiation in the prevalence of ASD is more attributable to the unique genetic architecture of the X chromosome hemizygosity (where alleles for males are expressed unbuffered by a second X chromosome), rather than female effects in autosomes. This approach provides a quantitative framework for how allele frequency, effect size, and model influence risk for disease and applies to sex-specific genetic models and analysis of complex diseases.

Exploring X-linked Contributions to Sex-Specific Prevalence Differences in Autism Spectrum Disorder (ASD)

By

Se Jun Park

David J. Cutler
Adviser

A thesis submitted to the Faculty of
Emory College of Arts and Sciences of Emory University
in partial fulfillment of the requirements of the degree of
Bachelor of Science with Honors

Department of Biology

2025

Acknowledgements

I would like to express my deepest gratitude to my advisor, Dr. David Cutler, for his invaluable guidance, insightful feedback, and unwavering support throughout the course of this project. His mentorship has been fundamental to both the direction and depth of this thesis.

I would also like to thank my committee members, Dr. Michal Arbilly and Dr. Hojin Kim, for their thoughtful input, encouragement, and for dedicating their time and expertise to review my work.

Contents

Chapter 1

Introduction to ASD and Allele Frequency Patterns

Sex differences in disease prevalence span a wide array of conditions, including autoimmune disorders, neurodevelopmental conditions, and psychiatric diseases (Hanamsagar and Bilbo 2015). These differences underscore the importance of investigating genetic factors that influence disease risk differently between males and females(Huang et al. 2023). Understanding these disparities is important for personalized medicine and public health to tailor treatments by taking into consideration the various sex-specific genetic and environmental factors (Shah et al. 2024). Among neurodevelopmental conditions, Autism Spectrum Disorder (ASD) has a great sex ratio disparity in prevalence (Werling and Geschwind 2013). Although the exact genetic mechanisms for this difference are yet to be studied, ASD does constitute a useful model in researching sex-based genetic contributions to disease susceptibility.

In fact, genetic contributions to the risk for complex diseases generally involve both autosomal and sex-linked chromosomes, each contributing in distinct ways to sex-based differences in disease prevalence (Martin et al., 2021). With the patient sample data obtained from Satterstrom et al. (2020), we investigated the genetic background of Autism Spectrum Disorder (ASD) by comparing the frequency of some autosomal variants of genes

associated with ASD, between cases (autistic patients) and controls (control subjects) in our study. Figure 1 presents the allele frequency distributions of male and female ASD patients, specifically for protein-truncating variants (PTV) and synonymous variants (SYN), providing evidence of potential sex differences in autism-associated gene networks for autosomal genes.

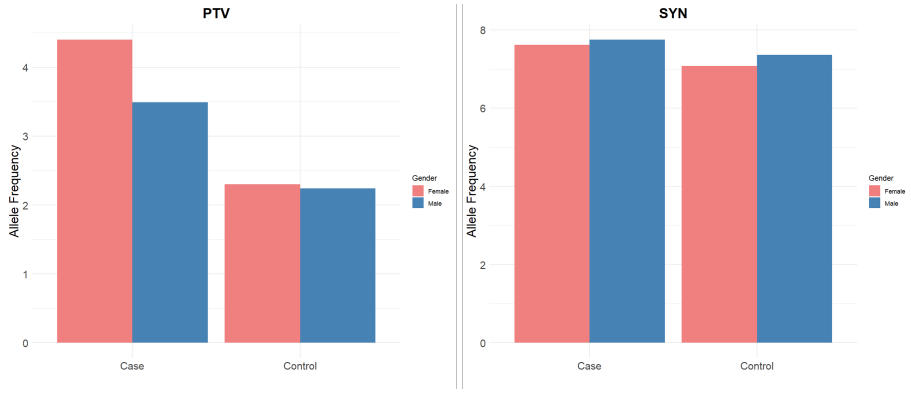


Figure 1.1: Allele frequency comparison in protein-truncating variants (PTVs) and synonymous variants (SYN) for male and female ASD cases and controls.

Autism Spectrum Disorder (ASD) is well known to be a male-biased disease, showing higher prevalence in males compared to females. However, through Figure 1, the risk alleles that are predicted to enhance disease risk, it was possible to observe a counterintuitive trend: these risk alleles, although enriched in cases compared to controls, occur at a higher frequency in females compared to males. If these alleles were causally implicated in the higher risk of ASD in males, they would be expected to be more common in affected males. Here, this paradox suggests that autosomal variants are not likely to be a significant factor in the sex differences seen in ASD. To answer this paradox, we will dive into the quantitative genetic models that can help us understand why these risk alleles are more prevalent in females if the disease is more prevalent among males.

Chapter 2

Introduction to the Liability Model

Using autosomal genes alone to explain the male-biased prevalence of ASD has its limitations, so we employ the liability model, a model that has been widely applied to dichotomous disease, to decipher these complex genetic processes. Liability is normally distributed in the population with a threshold that determines disease status: above the threshold, affected; below the threshold, unaffected.

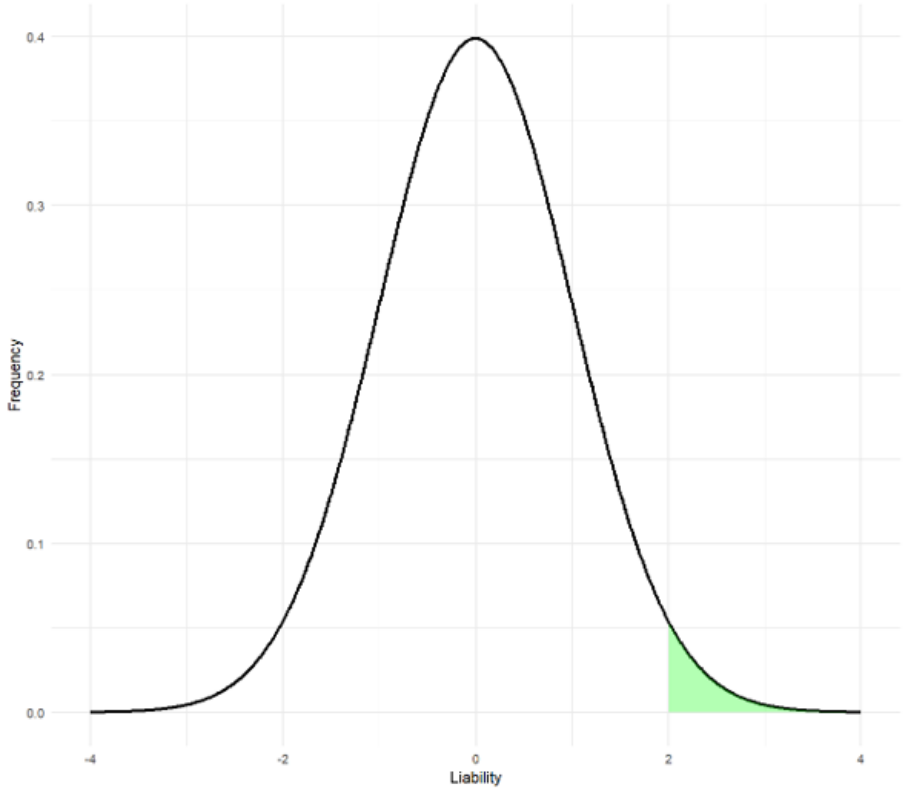


Figure 2.1: Liability threshold model showing risk for disease on a standard normal distribution.

According to Figure 2.1, this threshold model describes dichotomous diseases such as ASD extremely well because liability is considered a quantitative trait with a mean of 0 and variance of 1, and disease status depends only on whether liability exceeds a critical threshold. Thus, it suggests that the foundational principle of the liability model, expands the scope to be investigated based on sex-specific threshold values.

This liability model not only illustrates disease risk but provides a way to quantify the operationalizing of disease prevalence and threshold. With this model, we can to provide a definition of prevalence, which is the proportion of the subjects in a group that exhibit the disease. The threshold value on the liability scale, that is, t, was chosen to be equal to the observed prevalence, ψ, the threshold is based on the proportion of the population that has liability greater than t. This restates the threshold concept shown in Figure 2.1 above, where prevalence is linked as the precipitating factor for the threshold, where

people at a threshold of t or above are grouped as diseased. Algebraically, prevalence is the area under the liability distributions from t onward and can be expressed as:

$$\Pr[G \mid D] \Pr[D] = \Pr[D \mid G] \Pr[G]$$

where $\Phi(x)$ represents the standard normal probability density function, $\Phi(x)$ represents the cumulative distribution function, and $\Phi^{-1}(x)$ is the inverse of the cumulative function. With this in mind, it restates the threshold concept illustrated in Figure 2.1, connecting prevalence as the precipitating factor for the threshold at which people are grouped as diseased.

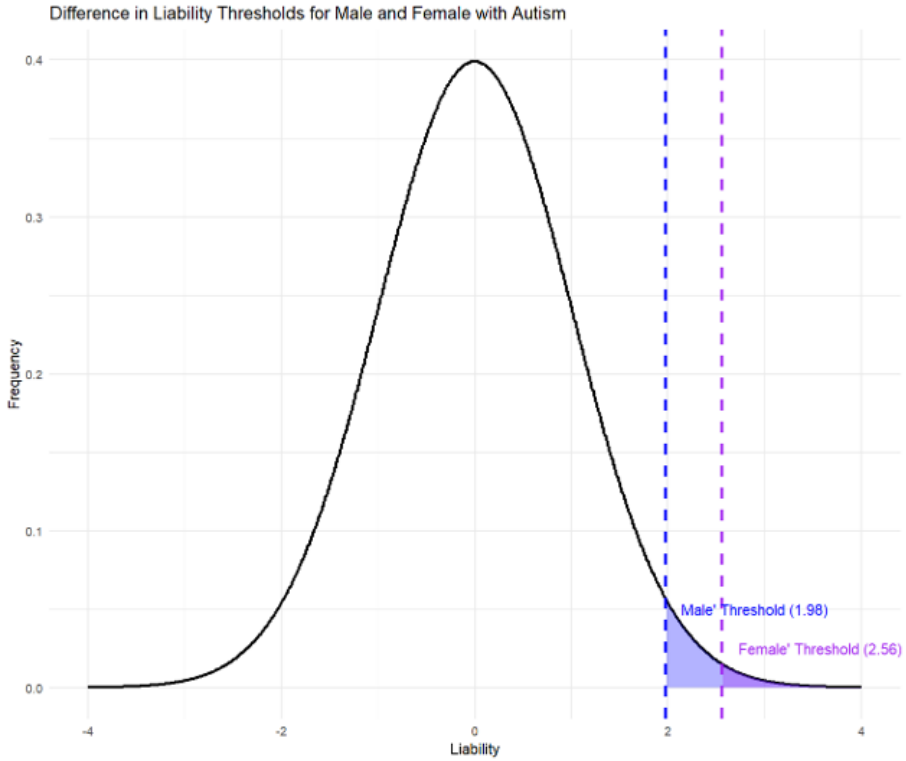


Figure 2.2: Comparison of male and female liability thresholds for autism, highlighting the lower threshold in males.

In fact, according to the Centers for Disease Control and Prevention (2014) data, approximately 1 in 42 boys is diagnosed with ASD compared to 1 in 189 girls in the U.S. population. The apparent difference means there are different liability thresholds for

males and females. Based on the male prevalence of 0.0238 (1 in 42), the threshold is approximately 1.98. For females, using the prevalence of 0.0053 (1 in 189), the threshold is about 2.56.

When we apply this to the liability model in Figure 2.2, you can see the difference as two plainly delineated threshold lines. The liability model shows that males and females have the same mean liability (i.e., similar genetic risk) for ASD. However, the threshold difference accounts for more males exceeding the threshold and being affected by ASD, resulting in males having more prevalence of ASD and that is what we see in reality. This liability model applied to ASD is instructive because it helps us visualize how both genetic and environmental risk factors for ASD, which are sex-specific, can differentially influence ASD risk in the two sexes.

Furthermore, beyond sex-specific thresholds, we have also shown that liability is also influenced by genetic variation within each sex, which we can measure by newly using penetrance: the likelihood of being diseased when they have a specific genotype. Figure 2.3 illustrates how risk in males may differ in regards to a shared genetic liability (allele A0 and A1) on the same phenotypic liability scale. Here, we can see, that when we observe other alleles they may have different impacts as reflected by their distributions which reflect the genetics responsible for alleles A0 and A1.

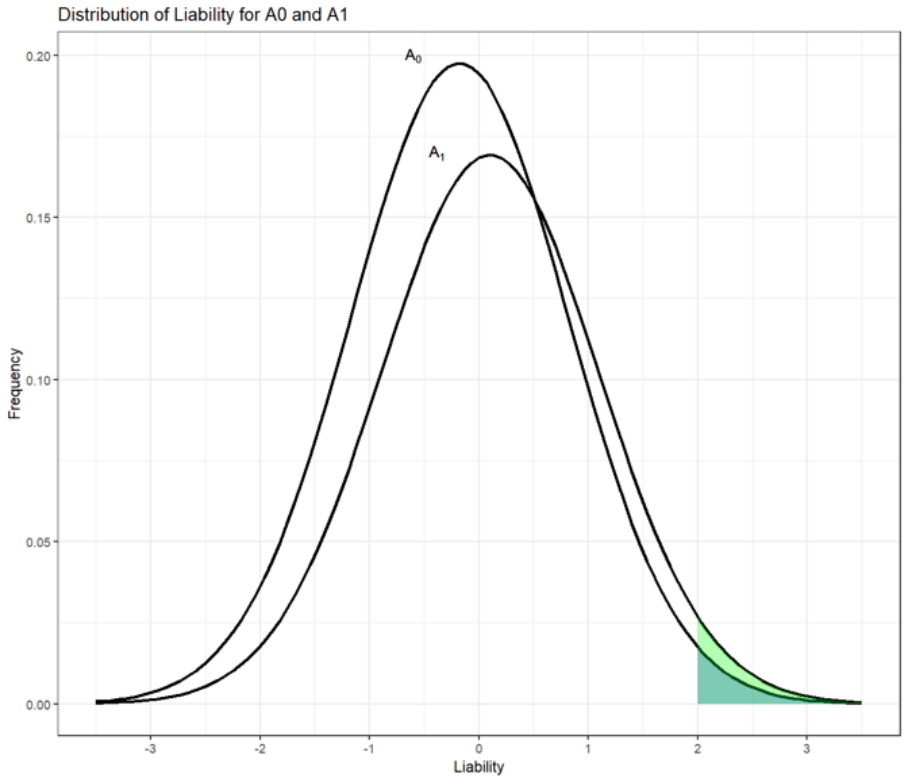


Figure 2.3: Alteration of liability distribution between A0 and A1 alleles, indicating higher penetrance for A1.

The penetrance differs between these alleles although their mean liability is equal because their shape or position functions are slightly different when put to the sex-specific threshold test. The area outside the threshold line for each allele represents the percentage of affected individuals and reveals that varying genetic forms can provide different penetrance values within the same sex group. The A1 allele carriers have more or less of a chance than those carrying A0 of exceeding the threshold based upon penetrance and how it operates.

The liability-threshold model illustrates that it accounts for sexual differences and also genetic difference using allelic examination in Figures 2.2 and 2.3 in conjunction. Figure 2.2 presents the fact that the liability thresholds exist at different levels across sexes which further upholds that ASD has increased incidence in males versus females. The particular genetic variants described in Figure 2.3 demonstrate how they have direct effects on levels

of ASD risk in the sex-specific subgroup while demonstrating the significance of genetic variation and sex-specific threshold levels in determining disease incidence.

The next section uses Bayesian methods to investigate average liability between affected male and female autism spectrum disorder individuals. This technique synthesizes allele frequencies, disease prevalence, and genetic risks to create an explicit picture of the average liability for females and males. The approach allows for calculation in terms of disease probability given a genotype, $\Pr(D \mid G)$, that ultimately determines the mean liability measure for females and males.

Bayes' Theorem provides the foundation for this analysis:

$$\Pr[G \mid D]\ \Pr[D] = \Pr[D \mid G]\ \Pr[G]$$

where: PrD: The prevalence of ASD, calculated from CDC data, provides the population probability of disease. $\Pr(G \mid D)$: The probability of having a specific genotype given that an individual has ASD, as we calculate from allele frequency data To proceed, we apply the law of total probability

$$\Pr[X] = \sum_{i=1}^{n} \Pr[X \mid Y_i]\ \Pr[Y_i],$$

to find the overall probability of a genotype Pr[G], as follows:

$$\Pr[G] = \Pr[G \mid D]\ \Pr[D]\ +\ \Pr[G \mid \neg D]\ \Pr[\neg D],$$

This leads to the calculation of $\Pr(D \mid G)$ which is the probability of ASD given a particular genotype and we can use this to estimate the influence of genetic factors on ASD risk for both males and females.

$$\Pr[D \mid G] = \frac{\Pr[G \mid D]\ \Pr[D]}{\Pr[G]},$$

Now that we have $\Pr(D \mid G)$, we can calculate the average liability, $\mathbb{E}[L \mid D]$ which is

the mean liability for individuals with the disease. This is done through the following equation: $\phi\big(T - \mathrm{E}[P \mid G]\big) = 1 - \Pr[D \mid G]$. Thus $\mathrm{E}[P \mid G] = T - \phi^{-1}\big(1 - \Pr[D \mid G]\big)$.

The $\mathbb{E}[L \mid D]$ value calculations between males and females provide crucial data to understand ASD liability variations between sexes because of genetic and environmental factors.

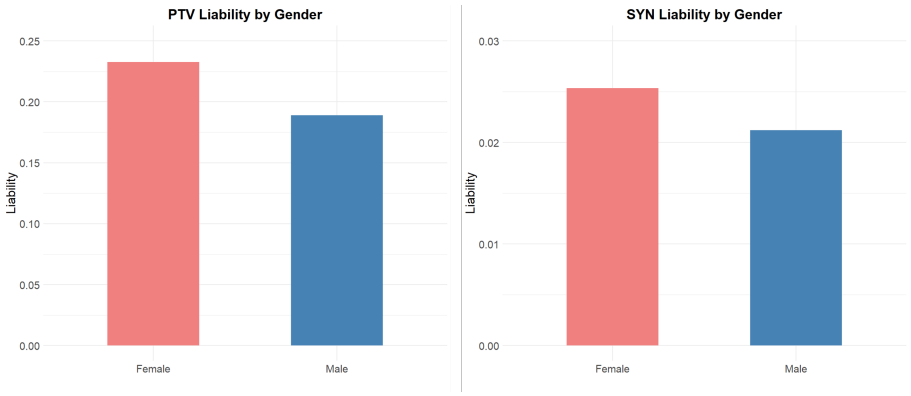


Figure 2.4: Sex-specific average liability values for protein-truncating (PTV) and synonymous (SYN) variants.

Figure 2.4 helps clarify the reason behind the higher allele frequency in females observed in Figure 2.1: their average liability for PTV and SYN variants is about the same for males and females, although the allele frequency has a difference. Also, the calculated prevalence of ASD showed a higher value in males than in females. These findings suggest that autosomal genes contribute little to no role and instead redirect our attention to the sex chromosomes, especially the X chromosome, as a potential major modifier of ASD's male-biased prevalence. The unique pattern of inheritance and expression for the X chromosome provides special insights into sex-linked genetic modifiers that may influence disease risk. We also emphasize that autosomal genes remain significant risk factors for people with ASD broadly, primarily through de novo and rare mutations disrupting neuronal and synaptic function. There, to fully understand sex differences in ASD prevalence, we need to quantitatively investigate additional genetic mechanisms in models. In particular, we will first analyze the liability associated with sex chromosomes through diverse modeling methods.

Chapter 3

Exploring Liability in Sex Chromosomes

Here this new model outlines a number of explanations for controlling sex-related prevalence difference disparities by changing thresholds and means or variances, or by utilizing a combination of these methods. However, in real life, prevalence differences in actual population samples derive primarily from simultaneous mean and variance divergences. The human sex difference in disease prevalence emerges as a consequence of male X-chromosome hemizygosity contrasting with female X-chromosome homozygosity or heterozygosity. When males express allelic effects they are enhanced since they lack a second identical chromosome to reduce the effect. Therefore, male liability distributions will have a higher mean and more spread values than females, elevating males' risk of exhibiting disease traits. From this model, it will provide additional information on sex-based effects by analyzing genetic profiles at the genotypic level.

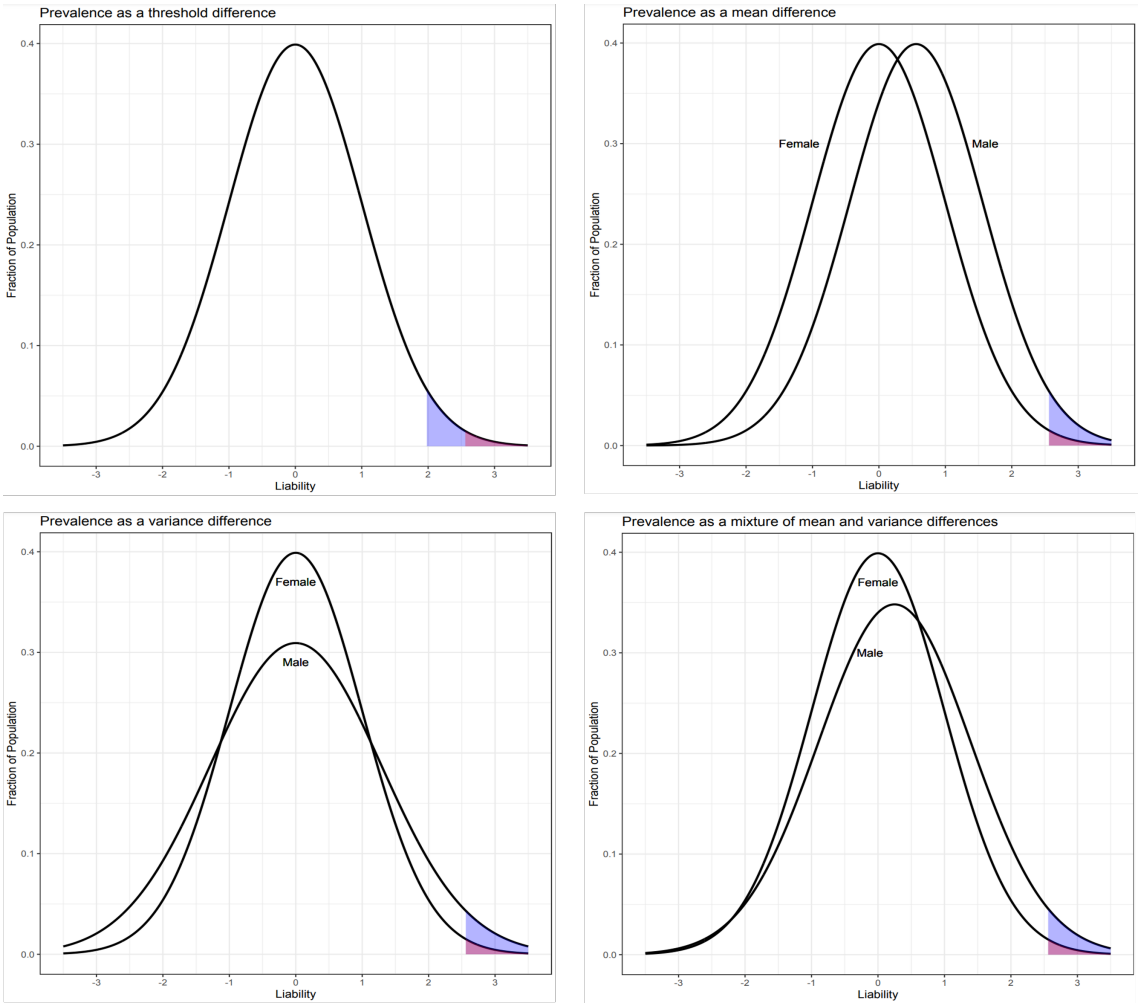


Figure 3.1: Models of prevalence differences driven by threshold, mean, variance, and combined liability shifts between sexes.

Chapter 4

Genetic Models and Their Impact on Liability

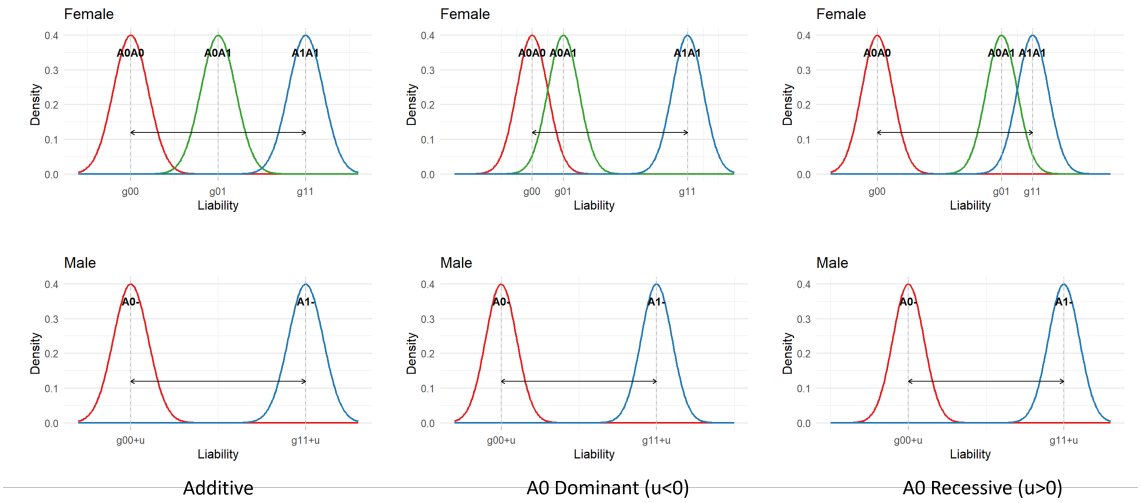


Figure 4.1: Comparison of additive, A0-dominant, and A0-recessive genetic models showing genotype-specific liability distributions by sex.

The following section provides three genetic models as examples of additive and recessive dominance inheritance with differences in means analysis for a fixed variance of one. Properties of additive models allow heterozygote females A0A1 to express traits at a value exactly halfway between homozygous phenotypes A0A0 and A1A1. The total distance

between mean X chromosome values is equal in both sexes because all baseline shifts are equally defined in males and females. When the phenotype of heterozygous female A0A1 moves towards the A0A0 type under the condition of dominance, the distance between the mean values in the population will still be the same. In males, the distribution is displaced to the left when u is negative, but the distance between allele effects is the same as for females. When A0 is recessive in its action the heterozygote is displaced in the direction of the A1A1 phenotype. These models show how biological mechanisms, for example, dosage compensation, keep average liability equivalent in sexes exhibiting a similar difference, even though males have greater variance because they are hemizygous. To account for sex-specific genetic effects the conceptual models need to be transformed into mathematical equations incorporating mean and variance parameters.

Chapter 5

Mathematical Framework for Sex-Based Liability Differences in Genetic Models

Now we can dive into math equations explaining the mean and variance values specifically, in additive, dominant, and total genetic versions. Here we consider a single diploid locus in Hardy-Weinberg equilibrium with two alleles, A0 and A1, where the frequency of A0 is p and the frequency of A1 is q=1-p. In general, we assume p is lager than q for most cases.

- α(Allelic Effect) Average phenotypic effect for a given allele, indicating its additive effect on the trait. It is the expected change in the phenotype when one allele is inherited.

- d (Dominance Deviation) Dominance effect is the departure of genotypic value from expectation based on additive effects only. Expression of interactions between pairs of genes is observed in estimates of dominance deviation for A0A0 or A1A1 homozygous genotypes that define the effect on phenotypic traits.

- g (Genotypic Effect) The genetic effect of a genotype minus the sum of the allelic effects. Phenotype expression results from synergistic interactions between both alleles of a single genotype.

In this model, α, d, and g are treated as random variables, each representing this locus's genetic, additive, and dominance effects. These equations and concepts are adapted from the foundational work in *The Quantitative Genetics of Human Disease: 1. Foundations* by D.J. Cutler et al. (2023)., which provides a foundational approach to modeling genetic contributions to complex traits.

Equation	**Female**	**Male**
$\mathrm{E}[a]$	$\begin{aligned}\mathrm{E}[a] &= \Pr[G = A_0A_0]\,(2\alpha_0) \\ &\quad + \Pr[G = A_0A_1]\,(\alpha_0 + \alpha_1) \\ &\quad + \Pr[G = A_1A_1]\,(2\alpha_1) \\ &= p^2(2\alpha_0) + 2pq(\alpha_0 + \alpha_1) + q^2(2\alpha_2) \\ &= 2\big(p\,\alpha_0 + q\,\alpha_1\big) \\ &= 0\end{aligned}$	$\begin{aligned}\mathrm{E}[a] &= p\,(2\alpha_0) + q\,(2\alpha_1) \\ &= 0\end{aligned}$
$\mathrm{Var}[a]$	$\begin{aligned}\mathrm{Var}[a] &= \mathrm{E}[a^2] - \big(\mathrm{E}[a]\big)^2 = \mathrm{E}[a^2] \\ &= p^2\,(2\alpha_0)^2 + 2pq\,(\alpha_0 + \alpha_1)^2 + q^2\,(2\alpha_2)^2 \\ &= 2\big(p\,\alpha_0^2 + q\,\alpha_1^2\big) \\ &= 2pq\,\big(\alpha_1 - \alpha_0\big)^2\end{aligned}$	$\begin{aligned}\mathrm{Var}[a] &= \mathrm{E}[a^2] - \big(\mathrm{E}[a]\big)^2 = \mathrm{E}[a^2] \\ &= p\,(2\alpha_0)^2 + q\,(2\alpha_1)^2 \\ &= 4\big(p\,\alpha_0^2 + q\,\alpha_1^2\big) \\ &= pq\,\big(2\alpha_1 - 2\alpha_0\big)^2\end{aligned}$
$\mathrm{E}[d]$	$\begin{aligned}\mathrm{E}[d] &= p^2\,d_{00} + 2pq\,d_{01} + q^2\,d_{11} \\ &= 0\end{aligned}$	$\begin{aligned}\mathrm{E}[d] &= p\,d_{00} + q\,d_{11} \\ &= \frac{p}{q}\,d_{00}\end{aligned}$
$\mathrm{Var}[d]$	$\begin{aligned}\mathrm{Var}[d] &= \mathrm{E}[d^2] - \big(\mathrm{E}[d]\big)^2 = \mathrm{E}[d^2] \\ &= p^2\,d_{00}^2 + 2pq\,d_{01}^2 + q^2\,d_{11}^2 \\ &= d_{00}\,d_{11}\end{aligned}$	$\begin{aligned}\mathrm{Var}[d] &= \mathrm{E}[d^2] - \big(\mathrm{E}[d]\big)^2 = \mathrm{E}[d^2] \\ &= pq\,\big(d_{11} - d_{00}\big)^2\end{aligned}$
$\mathrm{Var}[g]$	$\begin{aligned}\mathrm{Var}[g] &= \mathrm{Var}[a] + \mathrm{Var}[d] + 2\,\mathrm{Cov}[a,d] \\ &= \mathrm{Var}[a] + \mathrm{Var}[d] \\ &\quad (\text{since } \mathrm{Cov}[a,d] = 0)\end{aligned}$	$\begin{aligned}\mathrm{Var}[g] &= \mathrm{Var}[a] + \mathrm{Var}[d] + 2\,\mathrm{Cov}[a,d] \\ &= pq\,\big(g_{11} - g_{00}\big)^2\end{aligned}$

Table 1: Comparison of Female and Male Equations for $\mathrm{E}[a]$, $\mathrm{Var}[a]$, $\mathrm{E}[d]$, $\mathrm{Var}[d]$, and $\mathrm{Var}[g]$.

Relationship	**Equation**
Variance of a (Male vs. Female)	$\mathrm{Var}[a]_{\text{male}} = 2 \times \mathrm{Var}[a]_{\text{female}}$
Variance of d (Male vs. Female)	$\mathrm{Var}[d]_{\text{male}} = \frac{(p-q)^2}{pq} \times \mathrm{Var}[d]_{\text{female}}$
Variance of d (Female)	$\mathrm{Var}[d]_{\text{female}} = \big(\mathrm{E}[d]_{\text{male}}\big)^2$

Table 2: Relationships Between Male and Female Variance Terms.

In Table 1, the resulting equations between males and females have different outcomes because of X chromosome genetic characteristics. Both sexes have a predicted zero additive effect under the primary conditions of the model. Equilibrium conditions show that heterozygotic allele contributions neutralize one another so there will be no change to the mean additive effects.

A distinction begins to appear with the dominance factors. Females have a zero expected dominance effect because of their two X chromosomes, which cancel out allele interactions. In contrast to this, males have a nonzero dominance effect because of their single X chromosome that does not give them opposing allele dominance. Since males have only a single X chromosome, they have no second allele to oppose the dominance of the one that they inherit.

Also, the sex difference in the modification of additive traits occurs because the male population shows twice the amount of variance observed in females, i.e., $\mathrm{Var}[a]_{\mathrm{male}} = 2 \times \mathrm{Var}[a]_{\mathrm{female}}$. The absence of a second allele causes male-specific increased variation in additive genetic effects because they have only one X-linked allele. The relationship between male and female variance of dominance effects can be shown as a function of $\frac{(p-q)^2}{pq}$ because of allele frequency variation as well as a function of p and q values. Moreover, the patterns shared by $pq(B_1 - B_0)^2$ track the manner in which the Bernoulli process acts to delineate patterns of probabilistic allele expression in males via their X-chromosome hemizygotes.

Through these equations, the single X chromosome structure in combination with male hemizygosity shows how it produces equal additive effects between sexes but produces divergent dominant effects and increased variance in males. Here, it emphasizes understanding the study of sex-specific genetic effects on complex traits as the intrinsic X chromosome structure generates natural sex differences in variance but not variance resulting from extrinsic environment causes.

Additive Case:
We have g_{00}, g_{01}, g_{11} with

$$g_{01} = \frac{g_{00} + g_{11}}{2}, \quad g_{00} = -\frac{q}{p} g_{11}$$

Equation	**Female**	**Male**	**Relationship**
$\mathrm{E}[a]$	$\mathrm{E}[a]_f = 2(p\alpha_1 + q\alpha_2)$ $= 0$	$\mathrm{E}[a]_m = 0$	
$\mathrm{Var}[a]$	$\mathrm{Var}[a]_f = \frac{1}{2}pq(g_{11} - g_{00})^2$	$\mathrm{Var}[a]_m = pq(g_{11} - g_{00})^2$	$\frac{\mathrm{Var}[a]_f}{\mathrm{Var}[a]_m} = \frac{1}{2}$
$\mathrm{E}[d]$	$\mathrm{E}[d]_f = 0$	$\mathrm{E}[d]_m = 0$	
$\mathrm{Var}[d]$	$\mathrm{Var}[d]_f = 0$	$\mathrm{Var}[d]_m = 0$	
$\mathrm{Var}[g]$	$\mathrm{Var}[g]_f = \mathrm{Var}[a]_f + \mathrm{Var}[d]_f$ $= \frac{1}{2}pq(g_{11} - g_{00})^2$	$\mathrm{Var}[g]_m = pq(g_{11} - g_{00})^2$	

Table 3: Additive Case: Comparison of Female and Male Equations and Their Relationships

Applying the equations described in Table 1, we derived explicit equations for the additive model under the assumption that the value of g01 is half the sum of g00 and g11. Table 3 illustrates the equations for expected and variance values of the additive, dominance, and total genetic components both in males and females. It also provides a direct comparison between the two sexes from the ratio that is given. These relationships are critical to understanding how the additive model formed sex-specific liability parameters and will underlie graphical comparisons that describe how additive effects induce male-female distinctions in genetic liability.

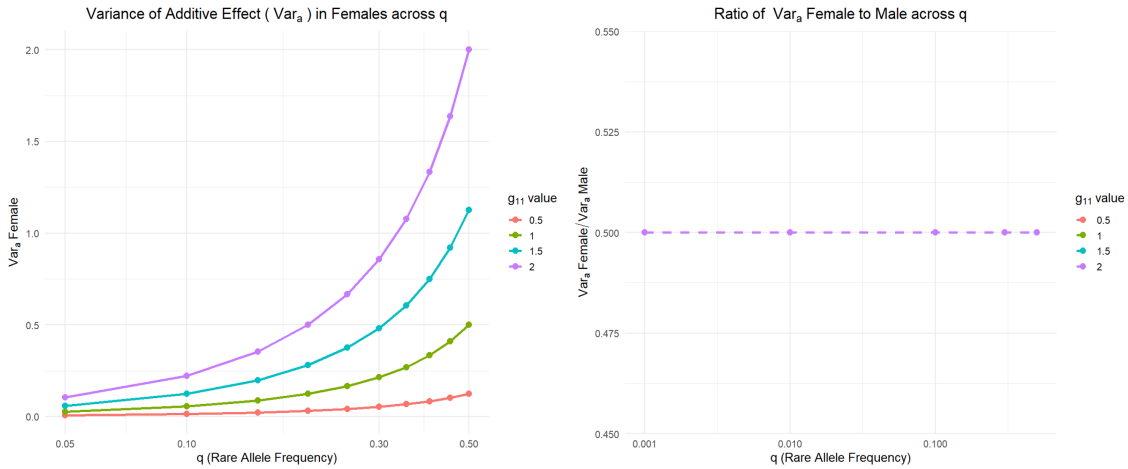


Figure 5.1: Variance of additive effect in females across rare allele frequencies and the female-to-male variance ratio under the additive model.

The figures generated for the additive model show that as the rare allele frequency increases, the additive variance increases as well. The female-to-male ratio remained constant at 0.5. One thing to note is that, since it is an additive model, dominance effects are not yet considered.

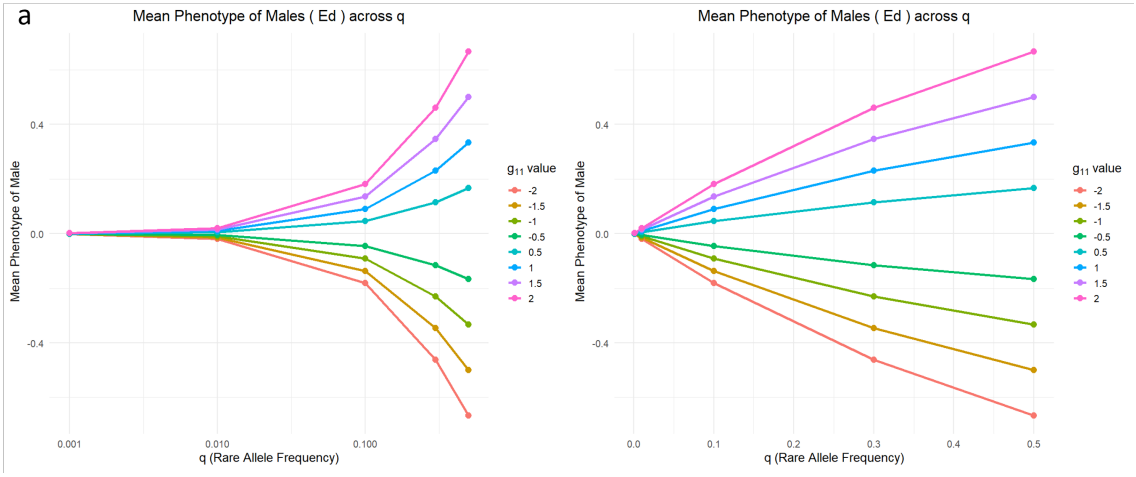
Fully Dominant Case:
We have g_{00}, g_{01}, g_{11} with $g_{00} = g_{01}$ and

$$g_{00} = -\frac{q^2}{p(1+q)}\, g_{11}$$

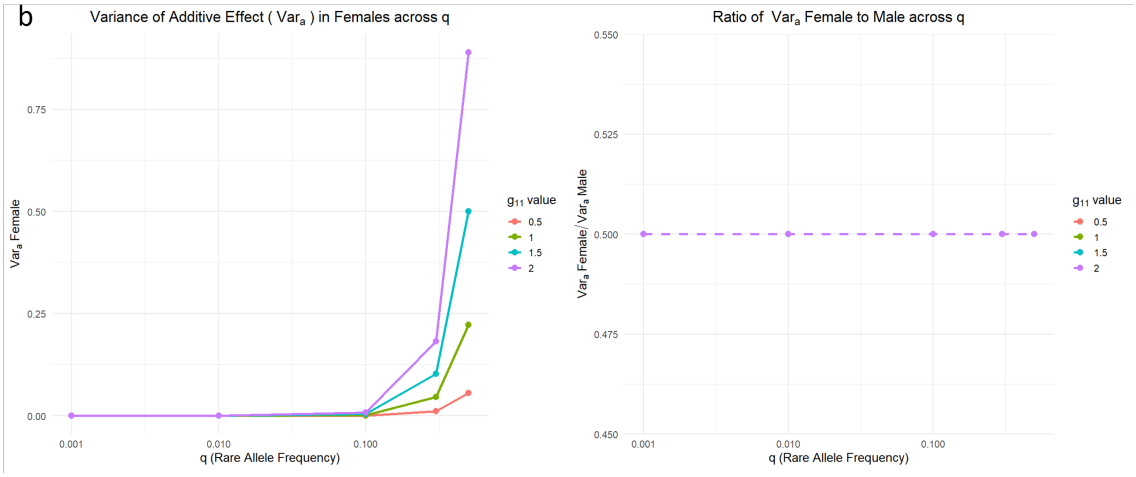
Equation	**Female**	**Male**	**Relationship**
$\mathrm{E}[a]$	$\mathrm{E}[a]_f = 2\big(p\,\alpha_1 + q\,\alpha_2\big)$ $= 0$	$\mathrm{E}[a]_m = p\,(2\alpha_0) \;+\; q\,(2\alpha_1)$ $= 0$	
$\mathrm{Var}[a]$	$\mathrm{Var}[a]_f = 2pq\,(\alpha_1 - \alpha_0)^2$ $= 2pq^3\,\big(g_{11} - g_{00}\big)^2$ $= \frac{2q^3}{p(1+q)}\, g_{11}^2$	$\mathrm{Var}[a]_m = pq\,\big(2\alpha_1 - 2\alpha_0\big)^2$ $= \frac{4q^3}{p(1+q)}\, g_{11}^2$	$\frac{\mathrm{Var}[a]_f}{\mathrm{Var}[a]_m} = \frac{1}{2}$
$\mathrm{E}[d]$	$\mathrm{E}[d]_f = p^2\, d_{00} \;+\; 2pq\, d_{01} \;+\; q^2\, d_{11}$ $= 0$	$\mathrm{E}[d]_m = p\, d_{00} \;+\; q\, d_{11}$ $= p\, g_{00} \;+\; q\, g_{11}$ $= \frac{q}{1+q}\, g_{11}$	
$\mathrm{Var}[d]$	$\mathrm{Var}[d]_f = d_{00}\, d_{11}$ $= \frac{p^2}{q^2}\, g_{00}^2 \;=\; \frac{q^2}{(1+q)^2}\, g_{11}^2$	$\mathrm{Var}[d]_m = pq\,\big(d_{11} - d_{00}\big)^2$ $= pq\,(p-q)^2\,\big(g_{11} - g_{00}\big)^2$ $= \frac{q\,(p-q)^2}{p(1+q)^2}\, g_{11}^2$	$\frac{\mathrm{Var}[d]_f}{\mathrm{Var}[d]_m} = \frac{pq}{(p-q)^2}$
$\mathrm{Var}[g]$	$\mathrm{Var}[g]_f = \mathrm{Var}[a]_f + \mathrm{Var}[d]_f$ $= q^2(1-q)^2\big(g_{11} - g_{00}\big)^2$ $= \frac{q^2}{p(1+q)}\, g_{11}^2$	$\mathrm{Var}[g]_m = pq\,\big(g_{11} - g_{00}\big)^2$ $= \frac{q}{p(1+q)^2}\, g_{11}^2$	$\frac{\mathrm{Var}[g]_f}{\mathrm{Var}[g]_m} = q(1+q)$
$\frac{\mathrm{Var}[a]}{\mathrm{Var}[g]}$	$\frac{\mathrm{Var}[a]_f}{\mathrm{Var}[g]_f} = \frac{2q}{1+q}$	$\frac{\mathrm{Var}[a]_m}{\mathrm{Var}[g]_m} = 4q^2$	$\frac{\left(\frac{\mathrm{Var}[a]}{\mathrm{Var}[g]}\right)_f}{\left(\frac{\mathrm{Var}[a]}{\mathrm{Var}[g]}\right)_m} = \frac{1}{2q(1+q)}$
$\frac{\mathrm{Var}[d]}{\mathrm{Var}[g]}$	$\frac{\mathrm{Var}[d]_f}{\mathrm{Var}[g]_f} = \frac{p}{1+q}$	$\frac{\mathrm{Var}[d]_m}{\mathrm{Var}[g]_m} = (p-q)^2$	$\frac{\left(\frac{\mathrm{Var}[d]}{\mathrm{Var}[g]}\right)_f}{\left(\frac{\mathrm{Var}[d]}{\mathrm{Var}[g]}\right)_m} = \frac{p}{(1+q)(p-q)^2}$

Table 4: Fully Dominant Case: Comparison of Female and Male Equations and Their Relationships

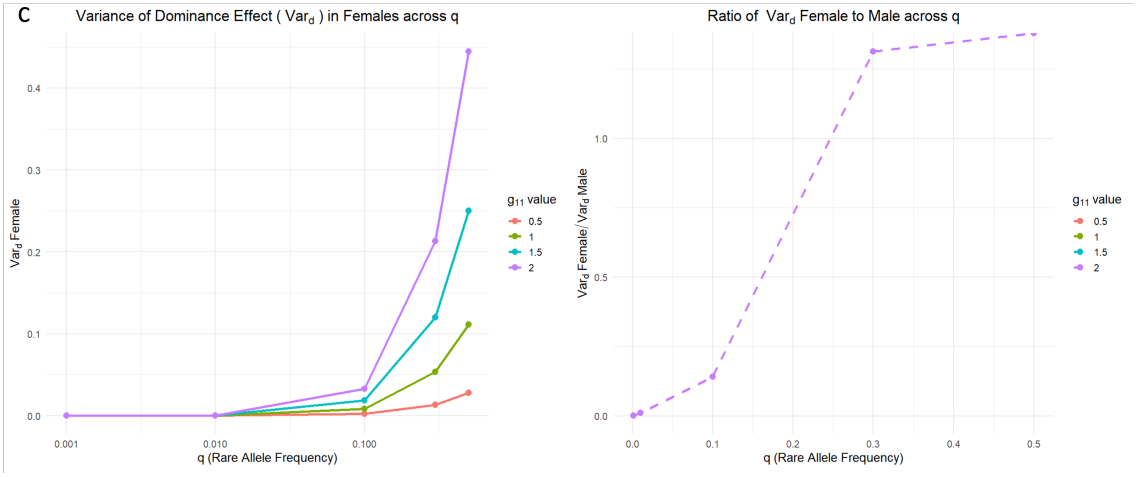
Assuming g00=g01 to denote the dominance of one allele over the other, the equations of the fully dominant model can be found in Table 4. Similar to Table 3, this table offers simplified expressions and enables graphical analyses, thus providing insight into the extent of full dominance in modulating sex differences in genetic variances.



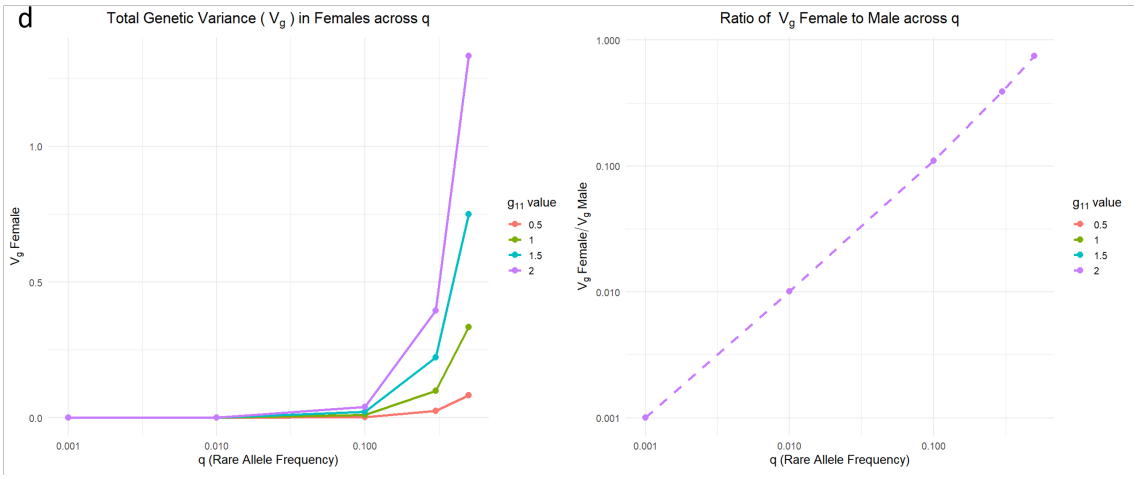
(a) Mean phenotype of males across rare allele frequencies under the fully dominant model.



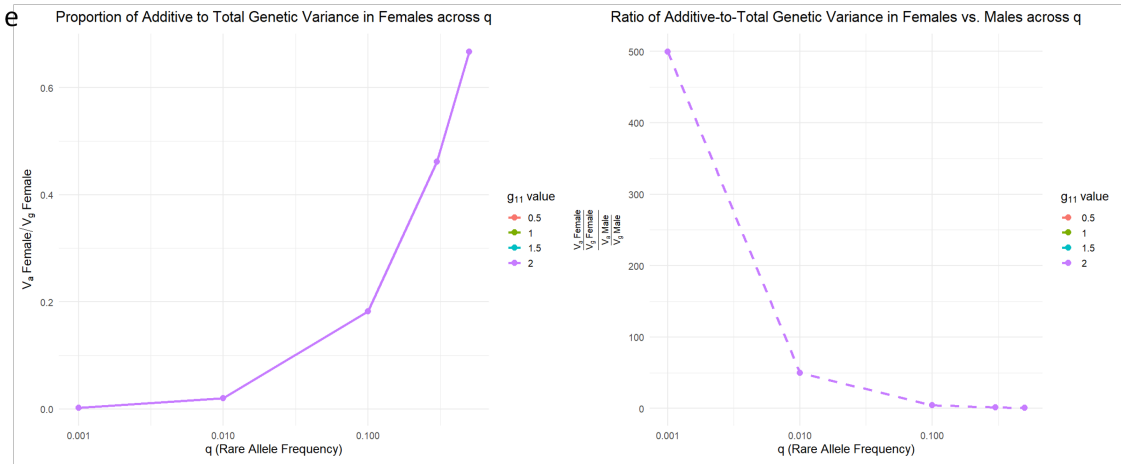
(b) Variance of additive effect in females across rare allele frequencies and the female-to-male variance ratio under the fully dominant model.



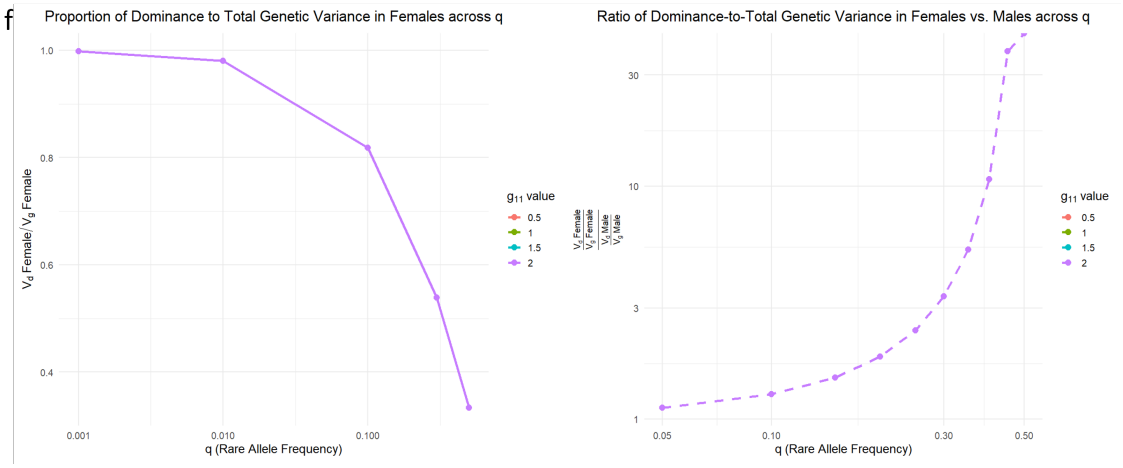
(c) Variance of dominance effect in females across rare allele frequencies and the female-to-male variance ratio under the fully dominant model.



(a) Total genetic variance in females across rare allele frequencies and the female-to-male variance ratio under the fully dominant model.



(b) Additive-to-total genetic variance in females and the female-to-male ratio of this proportion across rare allele frequencies under the fully dominant model.



(c) Dominance-to-total genetic variance in females and the female-to-male ratio of this proportion across rare allele frequencies under the fully dominant model.

Figure 5.3: Graphs under the fully dominant model illustrating genotype-specific trends across rare allele frequencies.

Data plotted in Figure 5.3(a) illustrate that an increase in the frequency (q) of the rare allele is associated with an increase in the mean phenotype of male subjects. This is a reflection of the increased likelihood of males having the dominant allele, whose influence is much greater in hemizygous male individuals. Figure 5.3(b) illustrates that additive genetic variance in females is extremely low at very low values of q but starts to increase gradually when q is more than approximately 0.01. This agrees with the fact that a very rare dominant allele has a very minor contribution to additive variance in females at the onset. Within this range, the ratio of additive female-to-male variance is always fixed at 0.5, in agreement with theoretical expectations; because of the hemizygosity of males, their additive variance must be twice that of females.

As illustrated in Figure 5.3(c), the variance of male dominance is higher than the variance of female dominance when the allele is rare because of the unmasked effect of one copy of the X-linked allele. However, as q approaches unity, female dominance variance increases sharply, eventually covering or even surpassing the sex difference at high frequencies. Figure 5.3(d) presents the total genetic variance in females, which quite closely replicates the overall pattern of the additive and dominance components. The ratio of female-to-male has an almost linear rise at low values of q but is afflicted with an acceleration as the allele frequency rises.

Figure 5.3(a) indicates that for low values of q, the additive variance component is an insignificant fraction of the overall genetic variance among females, demonstrating that dominance variance is the principal component of genetic influences under such a scenario. Particularly, the ratio of these fractions in females and males increases significantly, demonstrating that males, with a single allele, are more vulnerable to the influence of dominance when the allele is infrequent. Finally, Figure 5.3(f) highlights the striking presence of dominance variance proportion among female subjects, which remains present strongly even at lower q values. However, relative comparison shows that, due to their hemizygosity, males show an even higher level of dominance-caused variation when the allele is very rare.

Overall, this fully dominant model (in g00=g01) illustrates significant dominance effects across both genders when the allele is infrequent. However, due to the fact that males possess merely one copy of the X-linked allele, they tend to display more pronounced variations in genetic variance on average.

Fully Recessive Case:
We have g_{00}, g_{01}, g_{11} with $g_{01} = g_{11}$ and

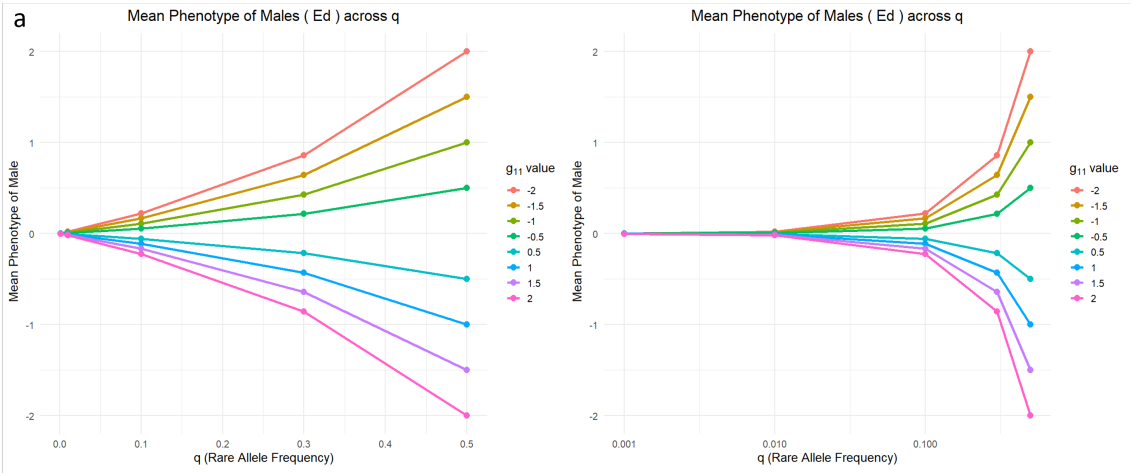
$$g_{00} = -\frac{q(p+1)}{p^2} g_{11}$$

Equation	**Female**	**Male**	**Relationship**
$\mathrm{E}[a]$	$\mathrm{E}[a]_f = 2(p\,\alpha_1 + q\,\alpha_2)$ $= 0$	$\mathrm{E}[a]_m = p(2\alpha_0) + q(2\alpha_1)$ $= 0$	
$\mathrm{Var}[a]$	$\mathrm{Var}[a]_f = 2pq(\alpha_1 - \alpha_0)^2$ $= 2p^3 q\,(g_{11} - g_{00})^2$ $= \frac{2q}{p} g_{11}^2$	$\mathrm{Var}[a]_m = pq(2\alpha_1 - 2\alpha_0)^2$ $= \frac{4q}{p} g_{11}^2$	$\frac{\mathrm{Var}[a]_f}{\mathrm{Var}[a]_m} = \frac{1}{2}$
$\mathrm{E}[d]$	$\mathrm{E}[d]_f = p^2 d_{00} + 2pqd_{01} + q^2 d_{11}$ $= 0$	$\mathrm{E}[d]_m = pd_{00} + qd_{11}$ $= pg_{00} + qg_{11}$ $= -\frac{q}{p} g_{11}$	
$\mathrm{Var}[d]$	$\mathrm{Var}[d]_f = d_{00} d_{11}$ $= \frac{q^2}{p^2} d_{11}^2$ $= \frac{q^2}{(1+q)^2} g_{11}^2$	$\mathrm{Var}[d]_m = pq(d_{11} - d_{00})^2$ $= q(p-q)^2 (g_{11} - g_{00})^2 / p^3$ $= \frac{q(p-q)^2}{p^3} g_{11}^2$	$\frac{\mathrm{Var}[d]_f}{\mathrm{Var}[d]_m} = \frac{pq}{(p-q)^2}$
$\mathrm{Var}[g]$	$\mathrm{Var}[g]_f = \mathrm{Var}[a]_f + \mathrm{Var}[d]_f$ $= p^2(1-p)^2 (g_{11} - g_{00})^2$ $= \frac{q(p+1)}{p^2} g_{11}^2$	$\mathrm{Var}[g]_m = pq(g_{11} - g_{00})^2$ $= \frac{q}{p^3} g_{11}^2$	$\frac{\mathrm{Var}[g]_f}{\mathrm{Var}[g]_m} = p(p+1)$
$\frac{\mathrm{Var}[a]}{\mathrm{Var}[g]}$	$\frac{\mathrm{Var}[a]_f}{\mathrm{Var}[g]_f} = \frac{2p}{1+p}$	$\frac{\mathrm{Var}[a]_m}{\mathrm{Var}[g]_m} = 4p^2$	$\frac{\left(\frac{\mathrm{Var}[a]}{\mathrm{Var}[g]}\right)_f}{\left(\frac{\mathrm{Var}[a]}{\mathrm{Var}[g]}\right)_m} = \frac{1}{2p(p+1)}$
$\frac{\mathrm{Var}[d]}{\mathrm{Var}[g]}$	$\frac{\mathrm{Var}[d]_f}{\mathrm{Var}[g]_f} = \frac{q}{1+p}$	$\frac{\mathrm{Var}[d]_m}{\mathrm{Var}[g]_m} = (p-q)^2$	$\frac{\left(\frac{\mathrm{Var}[d]}{\mathrm{Var}[g]}\right)_f}{\left(\frac{\mathrm{Var}[d]}{\mathrm{Var}[g]}\right)_m} = \frac{q}{(1+p)(p-q)^2}$

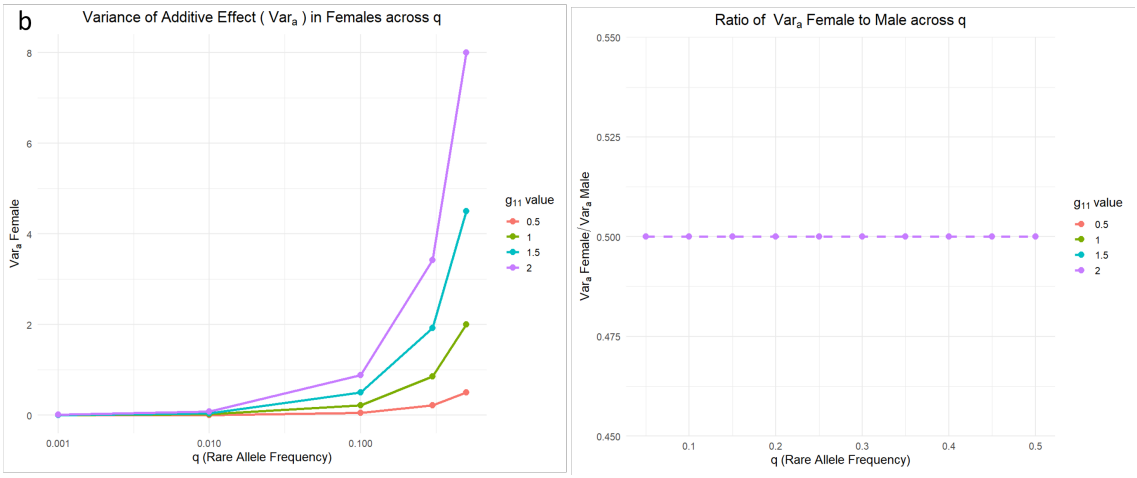
Table 5: Fully Recessive Case: Comparison of Female and Male Equations and Their Relationships

Table 5, presents equations for the fully recessive model, under the assumption g01=g11 to reflect the recessiveness of one allele over the other. Similar to Tables 3 and 4, it also contains simplified expressions, and supports graphical comparison, providing more information with which to consider how this fully recessive scenario might produce male-

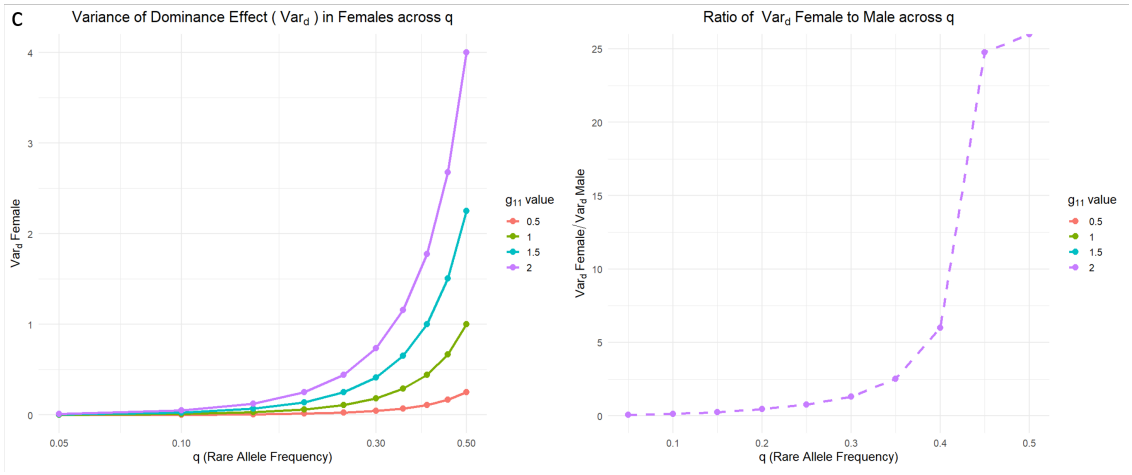
female differences in genetic liability.



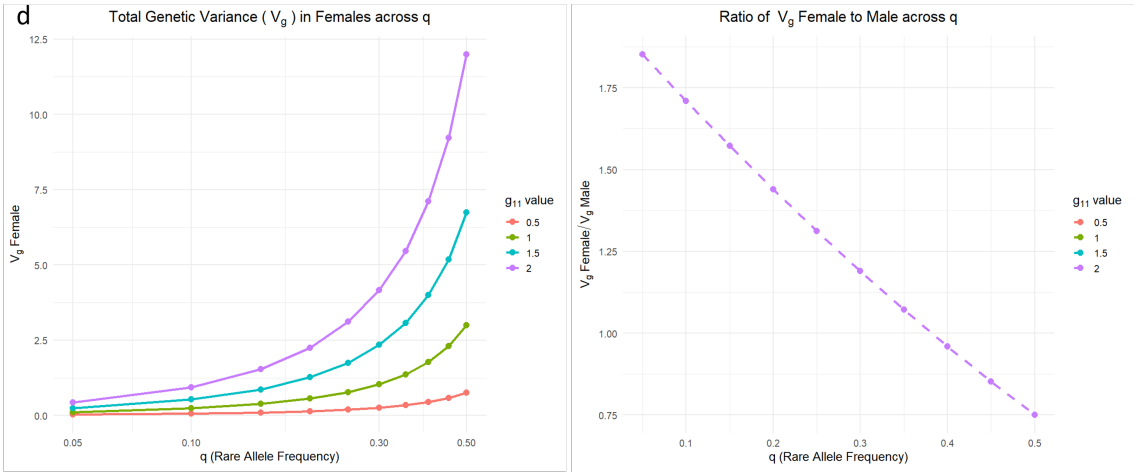
(a) Mean phenotype of males across rare allele frequencies under the fully recessive model.



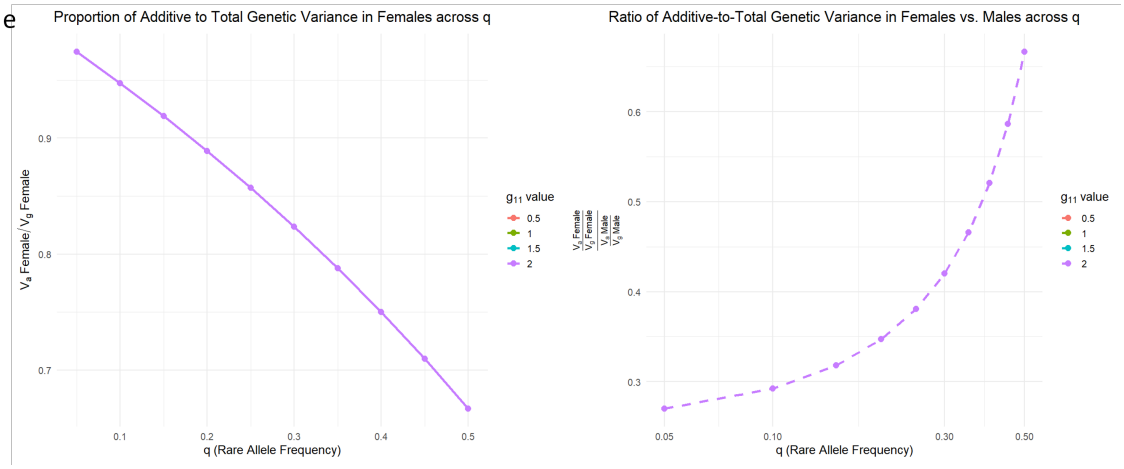
(b) Variance of additive effect in females across rare allele frequencies and the female-to-male variance ratio under the fully recessive model.



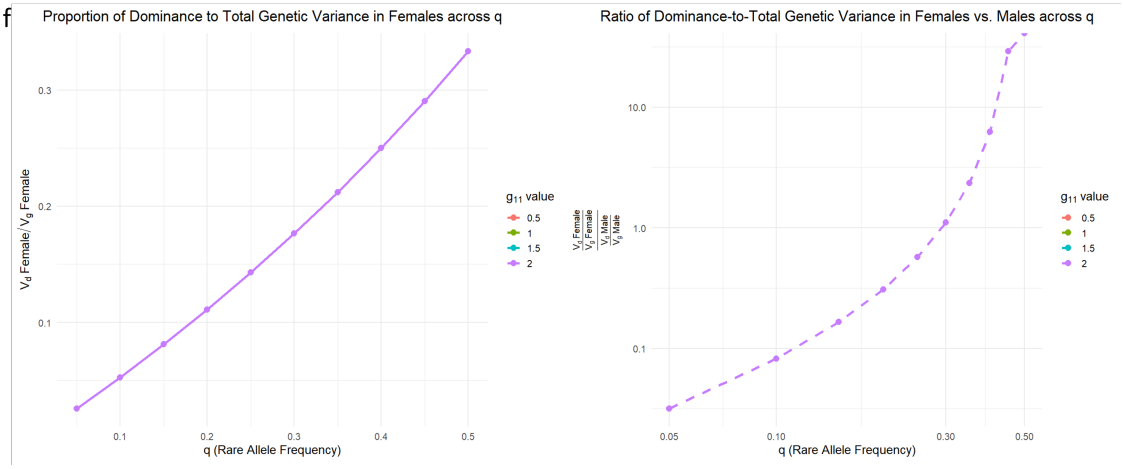
(c) Variance of dominance effect in females across rare allele frequencies and the female-to-male variance ratio under the fully recessive model.



(a) Total genetic variance in females across rare allele frequencies and the female-to-male variance ratio under the fully recessive model.



(b) Additive-to-total genetic variance in females and the female-to-male ratio of this proportion across rare allele frequencies under the fully recessive model.



(c) Dominance-to-total genetic variance in females and the female-to-male ratio of this proportion across rare allele frequencies under the fully recessive model.

Figure 5.5: Graphs under the fully recessive model illustrating genotype-specific trends across rare allele frequencies.

The recessive model generally provides the converse of the dominant model because the designation of an allele as dominant reverses the direction of the relationship between the genotype frequencies and phenotypic expression. In Figure 5.5(a), the average male phenotype increases as q values increase, but unlike Figure 5.3(a), it goes in the opposite direction because of the singular role played by g11 under recessive conditions. From Figures 5.5(b) and 5.5(c), we observe trends consistent with the dominant case: female-to-male additive variance ratio is fixed at 0.5, males (being hemizygous) have double the additive variance of females, and male dominance variance is higher than females for low q values. This again indicates that males, lacking a second X chromosome, can develop more pronounced dominance effects when the concerned allele is rare.

In Figure 5.5(d), the proportion of additive variance to total genetic variance is larger in females than males when q is small, indicating that under a fully recessive model, additive effects are more influential in the female population at first. Figure 5.5(e) also shows that with increasing q, the additive variance ratio of females to males will rise, showing that females' heterozygous and homozygous genotypes allow more scope for the expression of additive effects since the recessive allele is more common. Males either express or do not express the recessive trait, and there is less room for partial (additive) effects in hemizygous individuals.

Finally, Figure 5.5(f) shows female dominance variance low at low levels of q, while in males it is somewhat high again due to the uncomplicated effect of a single recessive allele on the X chromosome. Overall, while numerical structures are very dissimilar from the completely dominant case, hemizygosity remains to cause dominance effects larger in males at low levels of q, while variance components in females increase with rising q. These reversals of patterns from fully dominant to fully recessive models squarely rely on the base allele frequencies (p and q), testifying to the prime role that the distribution of alleles plays in building sex-differentiated liability profiles.

Chapter 6

Estimating the Number of X-Linked Risk Loci (n)

In the earlier sections, we calculated the components of mean and variance in both males and females given different models of genetics, e.g., fully dominant and fully recessive. These findings, regardless of the previously provided genotype effect, clearly explain how allele frequency and X-linkage lead to the liability distributions for each sex. Now that we have described these parameters, we can answer this question: Since we focused on a single locus, we now expand the analysis to the entire X chromosome and ask: how many X-linked loci (n) would be necessary to reconcile the observed differences in ASD prevalence between males and females?

To respond, we apply the quadratic formula to find both positive and negative roots in estimating n. This formula effectively combines our insights regarding mean and variance into a cohesive framework linking molecular genetic elements with population frequency at the organism level. The subsequent figures provide a clear visualization of how X-linked loci accumulate in different ways, influencing overall liability in males and females. Importantly, this analysis not only clarifies the fundamental genetic structure but also quantifies how X-linked factors contribute to the prevalence of diseases specific to each sex.

We wish to determine n, the total number of loci required to generate the observed prevalence difference in disease between males and females. In our analysis, the following relationships hold:

$$1 - \Phi\Big(t_f;\ nu_x,\ 1 + n\,\Delta V_{g,m}\Big) = \frac{1}{42},$$

$$1 - \Phi\Big(\frac{t_f - nu_x}{\sqrt{1 + n\,\Delta V_{g,m}}}\Big) = \frac{1}{42},$$

$$\frac{t_m^2}{t_f^2} = \frac{1}{1 + n\,\Delta V_{g,m}}\,.$$

Here,

- t_f is the threshold for females,
- t_m is the threshold for males,
- u_x is the mean value of x,
- $\Delta V_{g,m}$ is the change in genetic variance in males, and
- n is the total number of loci required.

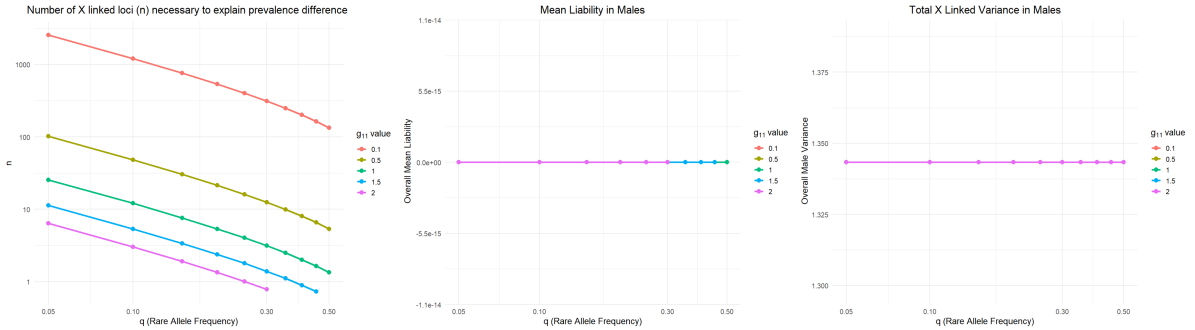
For the **additive model**, the number of loci is given by

$$n = \frac{\dfrac{t_f^2}{t_m^2} - 1}{\Delta V_{g,m}}\,.$$

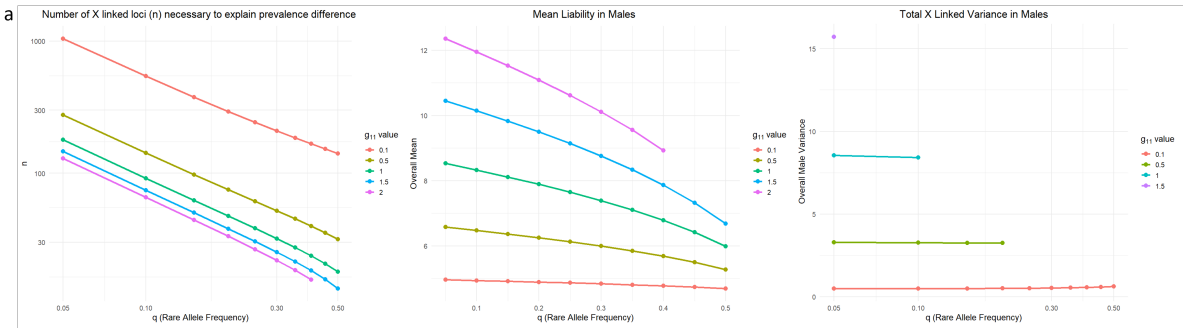
For the **fully dominant and fully recessive models**, solving the associated quadratic equation yields

$$n = \frac{2\,u_x\,t_f + t_m^2\,\Delta V_{g,m} \pm \sqrt{\Big(2\,u_x\,t_f + t_m^2\,\Delta V_{g,m}\Big)^2 - 4u_x^2\Big(t_f^2 - t_m^2\Big)}}{2u_x^2}\,.$$

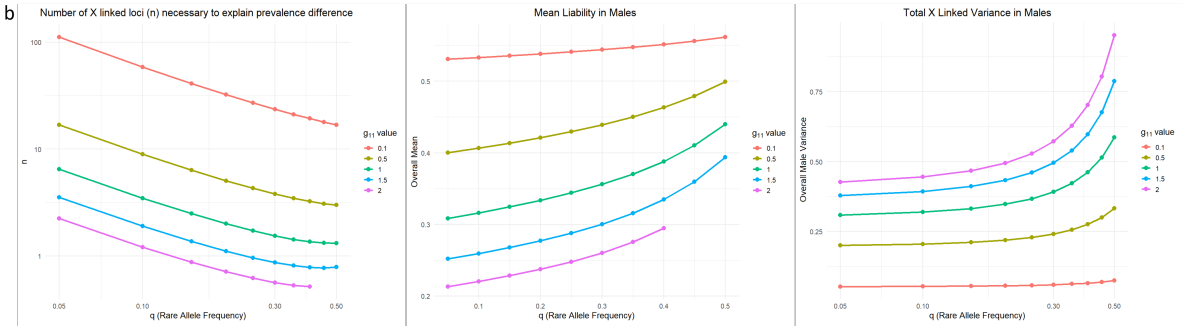
These expressions allow us to estimate the total number of loci n needed to produce the observed difference in disease prevalence between males and females.



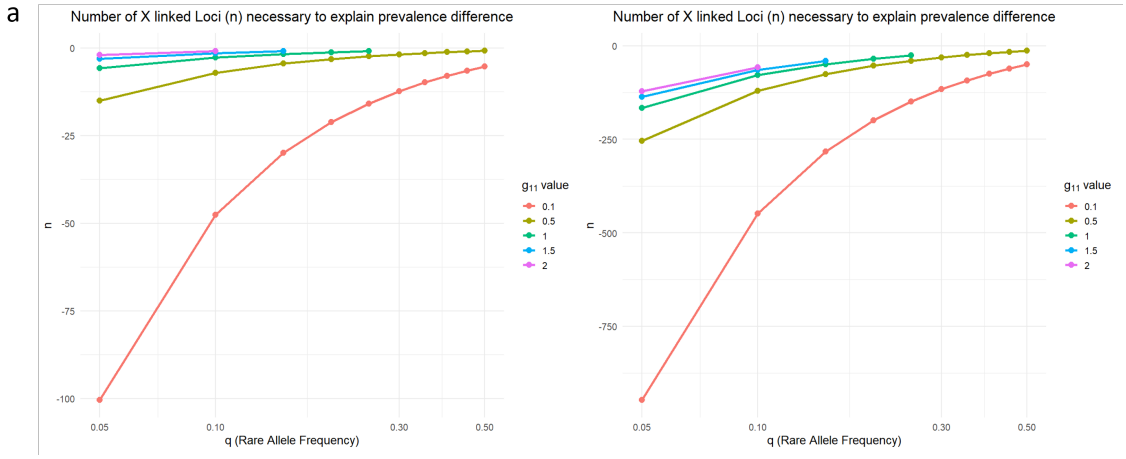
(a) Number of X linked loci (n) required to explain prevalence differences, mean liability in males, and total X-linked variance in males under the additive model.



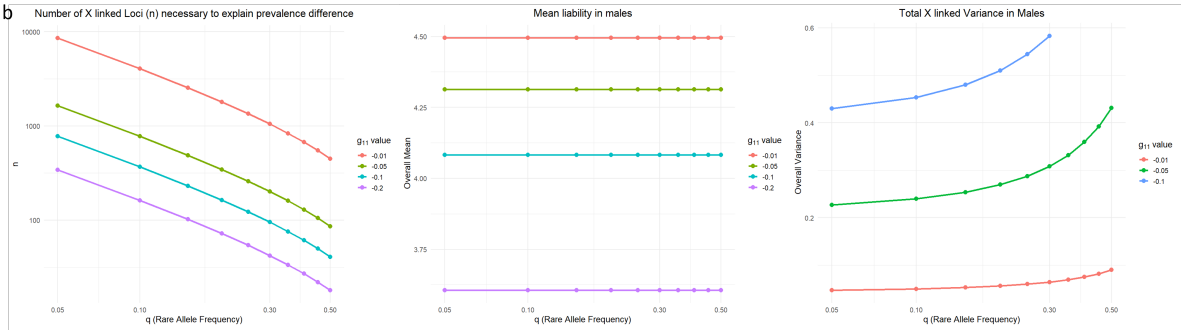
(b) Fully dominant model using the positive root of the quadratic solution to estimate the number of X-linked loci (n) required to account for sex-specific prevalence differences, along with the corresponding mean liability and total X-linked genetic variance in males.



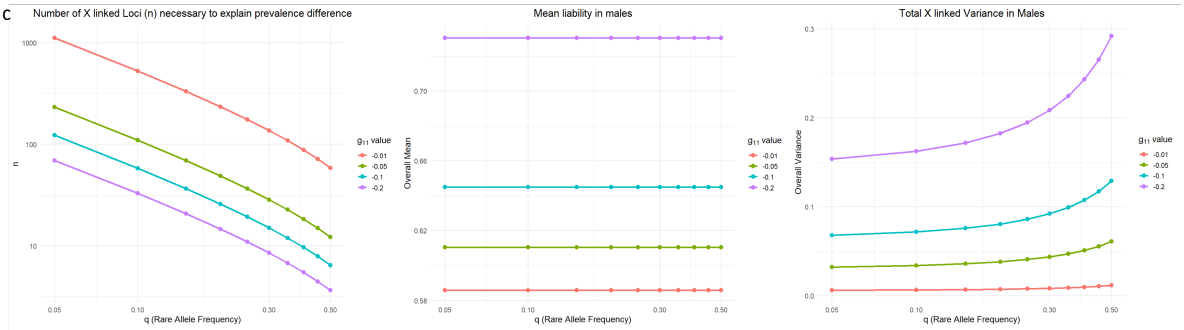
(c) Fully dominant model using the negative root of the quadratic solution to estimate the number of X-linked loci (n) required to account for sex-specific prevalence differences, along with the corresponding mean liability and total X-linked genetic variance in males.



(a) Fully recessive model using both the positive and negative root of the quadratic solution to estimate the number of X-linked loci (n) required to account for sex-specific prevalence differences when g11 is positive.



(b) Fully recessive model using the positive root of the quadratic solution to estimate the number of X-linked loci (n) required to account for sex-specific prevalence differences, along with the corresponding mean liability and total X-linked genetic variance in males when g11 is negative.



(c) Fully recessive model using the negative root of the quadratic solution to estimate the number of X-linked loci (n) required to account for sex-specific prevalence differences, along with the corresponding mean liability and total X-linked genetic variance in males when g11 is negative.

Figure 6.2: Graphs under different genetic models used to estimate the value of n, as well as the overall mean and variance

Chapter 7

Results

Under the additive model, our analysis shows that when there is no net mean difference between sexes, total variance is the predominant influence of differences in disease prevalence. In these cases, the additive variance among males is two times the additive variance for females. This implies that if females have about two-thirds the additive variance of males, then the disease prevalence differences observed must simply reflect variance differences. This scheme therefore provides some useful information on the nature of the interplay of multiple additive loci explaining sex-differentiated disease risk.

With respect to the fully dominant model, we can see the existence of a quadratic equation that has a negative and positive root. We will use the negative root because the mean liability must remain below the level required for disease expression. The number of loci associated with X-linkage needed to explain variations in prevalence under this model is very affected by effect size and allele frequency. An example would be a minor allele with a weak effect would need many loci to cause the difference in prevalence we see, while fewer loci with a larger effect would be needed. If many weak alleles are affected they mostly shift the mean liability, but if few strong alleles are affected the effect is mostly by variance inflation. Thus, these examples show that the same prevalence differences are contributed to either by few effective alleles or by many weak alleles, depending on whether that influences the mean or the variance.

In contrast, the fully recessive model presents a different pattern than the dominant model. For most models with normal parameters, we have negative estimates for the number of loci which suggests there is no plausible estimate for the number of loci unless we reverse the effect parameter (for example, g11 becomes negative). The explanation for this is that in a fully recessive model, the frequency of female heterozygotes (2pq) greatly exceeds the frequency of male hemizygotes (q). Hence, if the homozygous risk-enhancing genotype is positively influenced, the model suggests that females would have a correspondingly greater prevalence than males.

This also highlights that when g11 becomes negative, similar to what is observed in the fully dominant model, using the negative root of the quadratic solution allows us to estimate the number of X-linked loci(n) needed to account for the observed prevalence difference between males and females.

Overall, for an additive model, if the loci have large effect sizes, it could be as few as 1 to 10 loci. However, when individual effect sizes are small, thousands of loci might be necessary. In a fully dominant model, common alleles are dominant and protective, and the disease alleles are recessive, which is indicated by g11 having a positive value. Here if the effect sizes are large, it may take only a handful of loci (around 1-5) to be involved, whereas if they are relatively small, a hundred or more loci could be required. For the fully recessive model, where the rare allele is dominant (i.e, the heterozygote exhibits the same phenotype as the rare allele), the rare allele decreases the risk (serving a protective role) while the common allele increases the risk, evidenced by a negative g11 value. In the present case, the range of loci is 1 to 100 number of loci from moderately large effects to about thousands when the effects are small.

Combined, our results demonstrate that the number of loci necessary to account for any sex-specific prevalence is very sensitive to both the effect sizes and allele frequencies that create the genetic difference. In any case, whether the differences in rates arise solely from differences in means, variances, or both, we have given a framework to evaluate numerically and to consider, these various scenarios under this model for genetic clarification related

to any type of diversity. Importantly, the model indicated the relevance of additive and dominance effects underlying a trait associated with the X chromosomes.

Therefore, our findings are consistent with the likelihood that differences in the prevalence seen in this disorder are due to gene affecting the X chromosome via the unique biological phenomenon of male hemizygosity, rather than from multi-locus effects due to additive gene effects on the autosomes. The most in accord with our data interpretation is that there is a rare risk allele causing the disorder in males. Alternatively, if there were to be a common allele influencing the elevation risk, our model would require a high number of loci involved to account for the onset of differences in prevalence. Our results demonstrate that the prevalence of the disorder is higher in males than females consistent with the prediction that unbuffered expression of X-linked alleles increases liabilities to the disorder. These results suggest the disorder is genetically characterized by rare, strong effect alleles in males primarily, rather than by more ubiquitous autosomal contributions. Collectively these results aid in our understanding of sex-specific genetic effects on complex disorders and view the necessity for future research into X-linked contributory alleles in disease etiology.

Bibliography

Hanamsagar, R. and S. D. Bilbo (2015). "Sex Differences in Neurodevelopmental and Neurodegenerative Disorders: Focus on Microglial Function and Neuroinflammation during Development". In: *Brain, Behavior, and Immunity* 46, pp. 7–13. DOI: 10.1016/j.bbi.2015.02.006.

Huang, Y. et al. (2023). "Deciphering Genetic Causes for Sex Differences in Human Health through Drug Metabolism and Transporter Genes". In: *Nature Communications* 14.1, Article 1476. DOI: 10.1038/s41467-023-37038-1.

Shah, Y. R. et al. (2024). "Sex-based Disparities in Treatment and Healthcare Utilization in Patients with Ulcerative Colitis: A Systematic Review and Meta-analysis". In: *Journal Name* Volume.Issue, pages. DOI: 10.3390/jcm13247534.

Werling, D. M. and D. H. Geschwind (2013). "Sex Differences in Autism Spectrum Disorders". In: *Current Opinion in Neurology* 26.2, pp. 146–153. DOI: 10.1097/WCO.0b013e32835ee548.

Satterstrom, F. K. et al. (2020). "Large-Scale Exome Sequencing Study Implicates Both Developmental and Functional Changes in the Neurobiology of Autism". In: *Cell* 180.3, 568–584.e23. DOI: 10.1016/j.cell.2019.12.036.

Centers for Disease Control and Prevention (Mar. 27, 2014). *Autism Spectrum Disorder*. Accessed: [insert access date here]. URL: https://archive.cdc.gov/www_cdc_gov/media/releases/2014/p0327-autism-spectrum-disorder.html.