Spectrum of Disease Variation
Computational Genetics: Human Genetics 224
Jason Stein
May 30, 2008

Introduction:

Genome wide association studies (GWAS) are a recent advance in identifying genetic susceptibility to common and complex diseases[1], but it is not clear how much variation can be explained given a certain study size. Here we seek to write the total amount of genetic variation that can be explained given the parameters of a study.

Derivation:

Definition of Variables:

First, the variables defined in this derivation are listed in the table below

i refers to one SNP where each individual has I total SNPs that are in the genome
N is the total number of SNPs measured in the genome wide association study
M is the total number of unmeasured SNPS in the genome wide association study
g_i is the indicator of the genotype for one individual at one SNP i, which can have values $\{0,1\}$, where 1 is the risk allele and 0 is the non-risk allele
p_i is the minor allele frequency of a SNP and also equals $P(A_i)$
γ_i is the relative risk of the SNP for causing the phenotype of interest, or the probability of having the disease given the risk allele over having the disease given the absence of the risk allele $$\gamma_i = \frac{P(+\|A_i)}{P(+\|A_i^c)}$$
R is the baseline risk for having the phenotype that is independent of the relative risk contributed by the SNPs $$R = P(+\|A_1^c, \dots, A_I^c)$$ Where $A_1^c, \dots, A_I^c$ is the lack of all risk causing alleles.
F is the disease prevalence in the population and also equals $P(+)$
E is the percentage of disease variation that we are able to explain with GWAS
H^2 is the heritability of a disease
$\bar{P}$ is the average power of study

Calculation of Disease Prevalence:

The goal here is to write in short form how the disease prevalence, F, is a function of the minor allele frequencies of multiple alleles, p_i, as well as the relative risk, γ_i, and baseline risk, R. Let's take a simple example where there is only one risk allele in the genome for each person ($I = 1$).

There are two possible scenarios for the genotype.

Locus 1	**Probability**
$g = 1$	p_1
$g = 0$	$(1 - p_1)$

The disease prevalence, *F*, is then

$$F = p_1 P(+|A_1) + (1 - p_1) P(+|A_1^c)$$

$$F = p_1 P(+|A_1) + (1 - p_1) P(+|A_1)/\gamma_1$$

This can also be written as

$$F = R\gamma_1 p_1 + R(1 - p_1)$$

Where $R = P(+|A_1^c)$. Now we will expand the scenario to include two loci each with two possible alleles (I = 2). This scenario can be seen as

Locus 1	**Locus 2**	**Probability**
$g_1 = 1$	$g_2 = 0$	$p_1 (1 - p_2)$
$g_1 = 1$	$g_2 = 1$	$p_1 p_2$
$g_1 = 0$	$g_2 = 0$	$(1 - p_1) (1 - p_2)$
$g_1 = 0$	$g_2 = 1$	$p_2 (1 - p_1)$

The disease prevalence, *F*, is then

$$F = p_1 p_2 P(+|A_1, A_2) + p_1 (1 - p_2) P(+|A_1, A_2^c) + p_2 (1 - p_1) P(+|A_2, A_1^c) + (1 - p_1)(1 - p_2) P(+|A_1^c, A_2^c)$$

If A_1 and A_2 are independent (there is no LD between them), then $P(+|A_1, A_2) = P(+|A_1)P(+|A_2)$. Therefore the above formula can be written as

$$F = p_1 p_2 P(+|A_1)P(+|A_2) + p_1 (1 - p_2) P(+|A_1)P(+|A_2^c) + p_2 (1 - p_1) P(+|A_1^c)P(+|A_2) + (1 - p_1)(1 - p_2) P(+|A_1^c)P(+|A_2^c)$$

$$F = p_1 p_2 P(+|A_1)P(+|A_2) + p_1 (1 - p_2) P(+|A_1)P(+|A_2)/\gamma_2 + p_2 (1 - p_1) P(+|A_2)P(+|A_1)/\gamma_1 + (1 - p_1)(1 - p_2) P(+|A_1)P(+|A_2)/(\gamma_1 \gamma_2)$$

$$F = P(+|A_1)P(+|A_2)\left[p_1 p_2 + \frac{p_1 (1 - p_2)}{\gamma_2} + \frac{p_2 (1 - p_1)}{\gamma_1} + \frac{(1 - p_1)(1 - p_2)}{\gamma_1 \gamma_2}\right]$$

This can also be written as

$$F = R[\gamma_1 \gamma_2 p_1 p_2 + \gamma_1 p_1 (1 - p_2) + \gamma_2 p_2 (1 - p_1) + (1 - p_1)(1 - p_2)]$$

Where $R = P(+|A_1^c, A_2^c)$. To generalize this formula for I SNPs, we can say that

$$F = \prod_{i=1}^{I} P(+|A_i) \sum_{g_i=0}^{1} p_i g_i + \frac{(1 - p_i)(1 - g_i)}{\gamma_i}$$

Now let us assume there is a baseline risk, *R*, which is independent of having any measured risk genotype. This adjusts the above formula by

$$F = \prod_{i=1}^{I} R\gamma_i \sum_{g_i=0}^{1} p_i g_i + \frac{(1-p_i)(1-g_i)}{\gamma_i}$$

So the total disease prevalence is

$$F = R \prod_{i=1}^{I} \sum_{g_i=0}^{1} \gamma_i p_i g_i + (1-p_i)(1-g_i)$$

Amount of Disease Variation Explained

Now, assume we have *N* known SNPs that were measured, and *M* unknown SNPs that are unmeasured, where $M + N = I$. Also, assume that we know the disease prevalence in the population, *F*, through epidemiological studies. That means that the total amount of the genome that is explained by a study which measures *N* SNPs is a function of the minor allele frequency of each SNP, p_i, the disease prevalence, *F*, and the relative risk of each SNP, γ_i. Then, the percentage of disease variation that we can explain is

$$E(p_1, \dots, p_N, \gamma_1, \dots, \gamma_N, F) = \frac{R_N \prod_{i=1}^{N} \sum_{g_i=0}^{1} \gamma_i p_i g_i + (1-p_i)(1-g_i)}{F}$$

Let's take a simple example to demonstrate this point. Assume we measure two SNPs (*N* = 2) and that there are four total SNPs in the genome (*I* = 4), leaving two unmeasured SNPs (*M* = 2). According to the above derivation, the amount of genetic variation explained by this study would be

$$E = \frac{R_N[(1-p_1)+\gamma_1 p_1][(1-p_2)+\gamma_2 p_2]}{R_I[(1-p_1)+\gamma_1 p_1][(1-p_2)+\gamma_2 p_2][(1-p_3)+\gamma_3 p_3][(1-p_4)+\gamma_4 p_4]}$$

Here, the baseline risk are not the same in the numerator and the denominator as $R_N = P(+|A_1^c, A_2^c)$ and $R_I = P(+|A_1^c, A_2^c, A_3^c, A_4^c)$. However, we assume that if *N* is large enough $R_N = R_I$. That is that after you start having the absence of many risk alleles it will soon converge to the total baseline risk. One more important note is that $[(1-p_i)+\gamma_i p_i]$ will always be greater than or equal to one such that this quantity will always be a number between zero and one. Given this assumption, the above formula simplifies to

$$E = \frac{1}{[(1-p_3)+\gamma_3 p_3][(1-p_4)+\gamma_4 p_4]}$$

This is a very interesting result because it shows that the answer is independent of the measured SNPs. This allows the further quantification shown below. Another interesting result is that if the relative risks of these alleles are one ($\gamma_3 = 1, \gamma_4 = 1$) that the amount of the disease explained by the genetics is one. In other words, if these alleles do not cause risk for the disease then they are not considered.

Application

Application to find total unmeasured variation

In order to demonstrate an application of the above derivation, we can estimate the disease prevalence for a study with the following information where all disease alleles are assumed to be represented

$$\gamma = \{2.1, 2.3, 3.1, 2.8, 2.2, 3.0, 2.7, 2.5, 2.1, 2.1\}$$
$$p = \{0.15, 0.42, 0.32, 0.18, 0.34, 0.27, 0.08, 0.42, 0.08, 0.26\}$$
$$R = 0.001$$

With this information, $F = 0.0224$. Now we can remove some of the measured SNPs and determine the total amount of genetic information that is measured. For example, taking only the first 5 SNPs, how much does this study explain about the entire disease? Putting this into the above formula, we get that this study explains $E = 0.2506$ of the total genetic variation of the disease. This is implemented using the following Matlab code:

```
%Definition of Variables
gamma = [2.1,2.3,3.1,2.8,2.2,3.0,2.7,2.5,2.1,2.1];
p = [0.15,0.42,0.32,0.18,0.34,0.27,0.08,0.42,0.08,0.26];
R = 0.001;

%Calculation of disease prevalence
I = length(gamma);
F = R;
for i = 1:I
    summer = 0;
    for g = 0:1
        summer = summer + gamma(i)*p(i)*g + (1 - p(i))*(1-g);
    end
    F = F * summer;
end

%Calculation of percent explained by the study using the first 5 SNPs of
%the total risk alleles above
N = 5;
S = R;
for i = 1:N
    summer = 0;
    for g = 0:1
        summer = summer + gamma(i)*p(i)*g + (1 - p(i))*(1-g);
    end
    S = S * summer;
end

E = S/F;
```

A more useful application of this is to estimate the lower bound on the number of SNPs needed to completely describe a disease, given certain very specific constraints. Let's assume that we know there are *N* measured SNPs each with genotypes specified by $\gamma_1, \dots, \gamma_N > 2.0$ and $p_1, \dots, p_N > 0.05$ with $N = 10$. The disease prevalence, *F*, is also known through epidemiological studies. If there are *M* unmeasured SNPs, each with $\gamma_1, \dots, \gamma_M \leq 2.0$ and $p_1, \dots, p_M \leq 0.05$ how do we estimate the lower bound on *M*?

$$\gamma = \{2.1, 2.3, 3.1, 2.8, 2.2, 3.0, 2.7, 2.5, 2.1, 2.1\}$$
$$p = \{0.15, 0.42, 0.32, 0.18, 0.34, 0.27, 0.08, 0.42, 0.08, 0.26\}$$

$$R = 0.001$$
$$F = 0.10$$

By adding more SNPs with values of $p_i = 0.05$ and $\gamma_i = 2.0$, then an iterative process can be used to find when $E = 1$, or when the disease is completely explained by genetics. The value of E can be plotted as well to see how the addition of SNPs adds to the explanation of the disease. This implemented again using the following formula

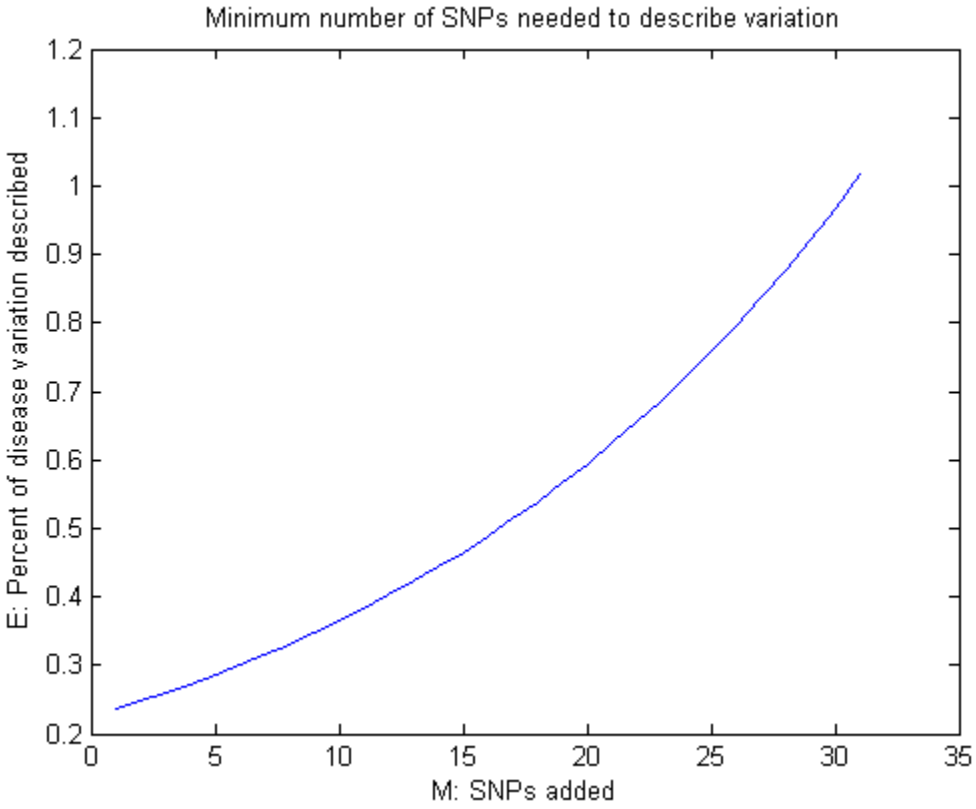
$$E(p_1, \ldots, p_N, \gamma_1, \ldots, \gamma_N, F) = \frac{R \prod_{i=1}^{N} \sum_{g_i=0}^{1} \gamma_i p_i g_i + (1 - p_i)(1 - g_i)}{F}$$

The minimum number of SNPs needed to describe the full variation of the disease can be calculated. The minimum number of other SNPs involved turned out to be 31 in this case. This was implemented in Matlab by the following code:

```
%Definition of Variables
gamma = [2.1,2.3,3.1,2.8,2.2,3.0,2.7,2.5,2.1,2.1];
p = [0.15,0.42,0.32,0.18,0.34,0.27,0.08,0.42,0.08,0.26];
R = 0.001;
F = 0.1;

%Calculation of percent explained by the study using the above SNPs
N = length(gamma);
S = R;
for i = 1:N
    summer = 0;
    for g = 0:1
        summer = summer + gamma(i)*p(i)*g + (1 - p(i))*(1-g);
    end
    S = S * summer;
end

E = S/F;

%Calculate the minimum number of SNPs that you are missing based on this
%information by adding SNPs with threshold values of gamma and p until 100
%percent E is reached
M = 1;
while E(end) < 1
    p(end + 1) = 0.05;
    gamma(end + 1) = 2;
    N = length(gamma);
    S = R;
    for i = 1:N
        summer = 0;
        for g = 0:1
            summer = summer + gamma(i)*p(i)*g + (1 - p(i))*(1-g);
        end
        S = S * summer;
    end
    E(M) = S/F;
    M = M + 1;
end
```

```
disp(['The minimum number of extra SNPs is: ' num2str(M-1)]);
plot(E);
xlabel('M: SNPs added');
ylabel('E: Percent of disease variation described');
title('Minimum number of SNPs needed to describe variation');
```

And is described by the graph



It is seen that the increase in explanation of the disease is not linear with the number of SNPs added. A minimum of 31 SNPs are needed to explain the full disease.

Heritability

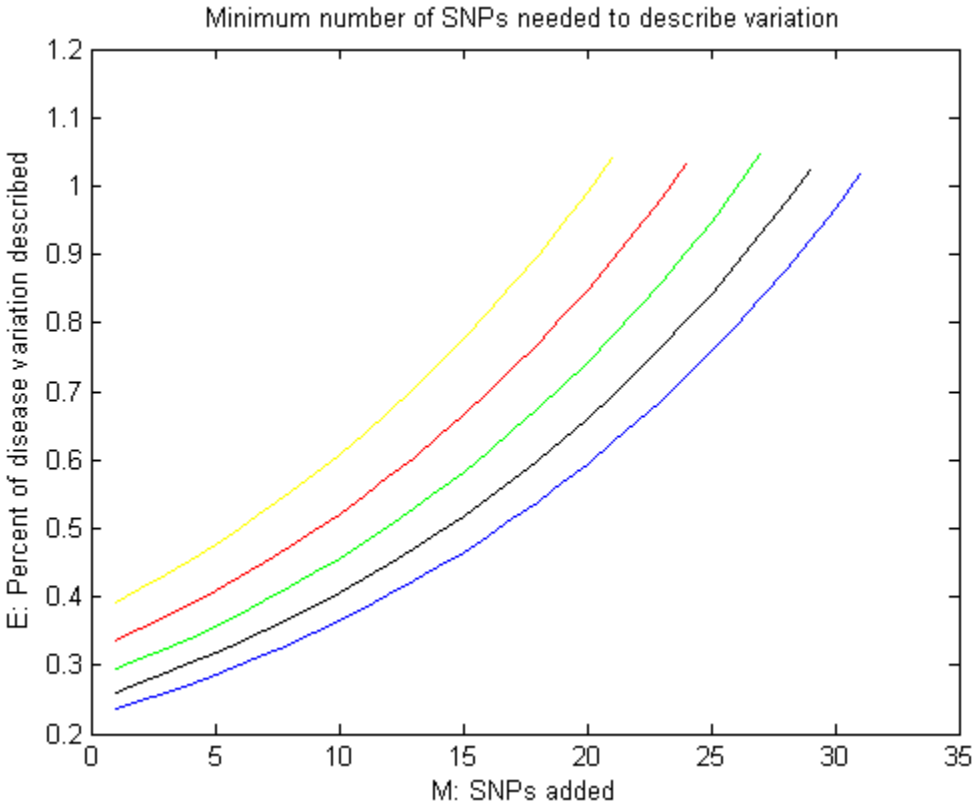
The reason for adding heritability

An implicit assumption in the above derivation and example is that the disease is entirely genetic in origin. In reality, most common disorders are not completely genetic but are a mix of genetics and environment. The quantification of the amount of disease variation explained by genetics is called heritability, H^2, and is derived from family studies and twin studies. The higher the heritability, the more that the genetic study should explain with the upper limit being the disease prevalence in the population. The lower the heritability, the less the genetic study should explain. This can be incorporated into our formula as

$$E(p_1, \dots, p_N, \gamma_1, \dots, \gamma_N, F, H^2) = \frac{R \prod_{i=1}^{N} \sum_{g_i=0}^{1} \gamma_i p_i g_i + (1 - p_i)(1 - g_i)}{H^2 F}$$

A continued application with heritability

Using the same situation as above, but with varying heritability values, it can be shown how heritability affects how much it is that we do not know. $H^2 = 1.0$ corresponds to a perfectly genetic disorder and is exactly the same as the previous situation where we did not take heritability into account. The lower the heritability, the less genetics can explain and therefore the less extra SNPs are needed to describe the disorder.



$H^2 = 0.6$: Yellow
$H^2 = 0.7$: Red
$H^2 = 0.8$: Green
$H^2 = 0.9$: Black
$H^2 = 1.0$: Blue

Power

The reason for adding power

The previous derivations assumed perfect power. Power gives the probability of detecting a significant allele associated with the disease given the allele frequencies, p_i, and the number of cases or controls in the sample, $\eta/2$. This means that all risk alleles only have a certain probability of detection so may have been missed. Decreased power will likely result in more alleles needed to explain the genetic variation.

Power Calculation

The power of the GWAS studies in which *N* SNPs are measured can be calculated given the allele frequencies and the number of subjects in a study. We know that for each SNP, the allele frequencies can be calculated from the relative risks, γ_i, and disease prevalence, *F*, as

$$P(A_i|+) = \frac{P(+|A_i)P(A_i)}{P(+)} = \frac{P(+|A_i)P(A_i)}{F} = \frac{\gamma_i P(A_i)}{P(A_i)(\gamma_i - 1) + 1}$$

And $P(A|-)$, where we cannot assume that *F* is small because we are concerned with common diseases, can be calculated as

$$P(-|A_i) = 1 - P(+|A_i) = 1 - \frac{\gamma_i F}{P(A_i)(\gamma_i - 1) + 1}$$

$$P(A_i|-) = \frac{P(-|A_i)P(A_i)}{P(-)} = \frac{P(A_i)\left(1 - \frac{\gamma_i F}{P(A_i)(\gamma_i - 1) + 1}\right)}{1 - F}$$

The non-centrality parameter for each SNP is calculated as

$$\lambda_i\sqrt{\eta} = \frac{P(A_i|+) - P(A_i|-)}{\sqrt{2P(A_i)(1 - P(A_i))}}$$

The average power of an entire study, $\bar{P}$, is the average of the power across all SNPs where there are *N* total SNPs measured and using a conservative Bonferroni correction.

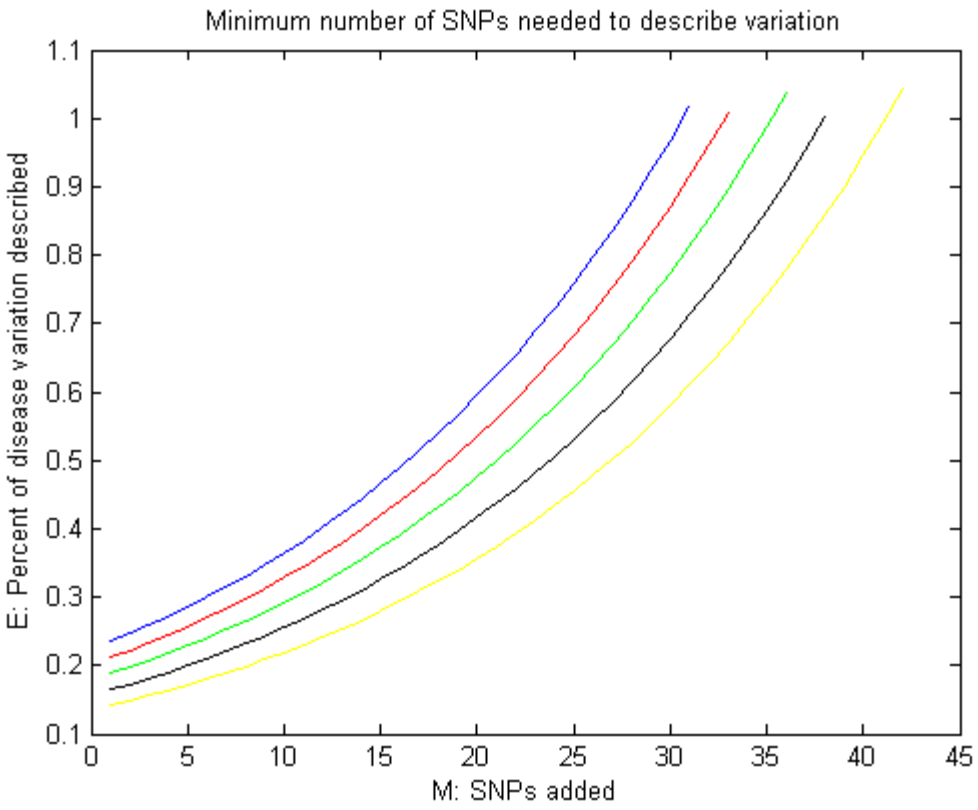
$$\bar{P} = \frac{1}{N}\sum_{i=1}^{N} 1 - \left(\Phi\left(\Phi^{-1}\left(1 - \frac{\alpha}{2N}\right) + \lambda_i\sqrt{\eta}\right) - \Phi\left(\Phi^{-1}\left(\frac{\alpha}{2N}\right) + \lambda_i\sqrt{\eta}\right)\right)$$

Power is then added to our above formula by

$$E(p_1, \ldots, p_N, \gamma_1, \ldots, \gamma_N, F, H^2, \alpha, \eta) = \frac{\bar{P}R\prod_{i=1}^{N}\sum_{g_i=0}^{1}\gamma_i p_i g_i + (1 - p_i)(1 - g_i)}{H^2 F}$$

A continued application with power

Using the same situation as above but varying power values and keeping heritability constant, we can establish a similar situation.



$\bar{P} = 1.0$: Blue
$\bar{P} = 0.9$: Red
$\bar{P} = 0.8$: Green
$\bar{P} = 0.7$: Black
$\bar{P} = 0.6$: Yellow

The blue line is the same as the original example. It can be seen that as the power of the study decreases, there is greater number of SNPs to be added.

References:

1. Hirschhorn, J.N. & Daly, M.J. Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet* **6**, 95-108(2005).