Nathaniel Dang
403221488
CS 124, Spring 2008

Project Documentation: Personalized Medicine

Motivation

There are many factors that must be taken into account when we calculate disease risk – including environmental and genetic factors, as well as occasional chance factors. If we are able to effectively approximate an individual's increased disease risk factor due to genetic variation, there are several actions we can take (assuming that other risk factors can be affected by such actions). These actions include more frequent and earlier screening for disease, pre-symptomatic medication aimed at reducing one or more risk factors, and focusing on reducing environmental risk factors, including making lifestyle changes. We can also use the information from the genetic study to study variations in more depth, to find correlated SNPs, and to further study the proteins encoded by the genes – which leads to a greater understanding of the overall genetic contribution to disease. There are several companies that offer a genetic "pre-screening" for disease risk, such as 23andMe and Navigenics, which take a similar approach.

Goals

We wish to be able to calculate an individual's risk for a certain disease given a baseline risk level, the presence or absence of several mutations, and the relative risks of the mutations. We also wish to be able to calculate the overall prevalence of a disease, as well the total genetic contribution to disease risk, given the minor allele frequencies and relative risks of mutations. Finally, we hope to study not only how to calculate the total genetic contribution to risk, but how *much* each factor affects this risk. For example, what has more of a genetic effect on overall risk – to have more SNPs of lower relative risk or probability, or to have fewer SNPs with a higher relative risk or probability?

Background

We wished to find how SNP(s) affected our disease risk. To simplify the problem, we made the assumption that SNPs are not correlated, and thus their effects on disease risk are completely independent.

- In the Single-SNP case, the probability of an individual having the disease without a mutation is given by $R = P(+|g_1=0)$, where R is our baseline level of risk.
- Then, the probability of having a disease given a mutation is: $P(+|g_1=1) = \gamma_1 R$
- Thus, the prevalence in the entire population is: $F=P_1\gamma_1 R + (1-P_1)R$
- Then, the genetic contribution can be found by dividing F by the baseline risk:
  $F/R=P_1 \gamma_1+(1-P_1)$
- In the Two-SNP case, the probability of an individual having the disease without a mutation is given by $P(+ \mid g_1=0, g_2=0) = R$, where R is our baseline level of risk.
- Then, the probability of having disease given a mutation at SNP_1 but not SNP_2 is:
  $P(+ \mid g_1=1,g_2=0) = \gamma_1 R$
- And the probability of having disease given a mutation at SNP_2 but not SNP_1 is:
  $P(+ \mid g_1=0,g_2=1) = \gamma_2 R$
- Thus, the prevalence in the entire population is:
  - $F =P_1P_2\, \gamma_1\, \gamma_2 R + P_1(1-P_2)\, \gamma_1\, R + P_2(1-P_1)\, \gamma_2 R + (1-P_1)(1-P_2)R$
- Then, the genetic contribution can be found by dividing by the baseline risk:
  - $F/R = P_1P_2\, \gamma_1\, \gamma_2 + P_1(1-P_2)\, \gamma_1 + P_2(1-P_1)\, \gamma_2 + (1-P_1)(1-P_2)$

However, we also wished to study how the presence of an unspecified *n* SNPs affected risk. When holding the minor allele frequencies p_i and relative risks γ_i constant across all SNPs, it can be shown that for *n* SNPs the genetic contribution to risk is given by:

- F/R=

$$\sum_{g\in\{0,1\}^n}\prod_{i=1}^{n}(p_i g_i+(1-p_i)(1-g_i))(\gamma_i)^{g_i}$$

Where g_i is the genotype for SNP_i. (Either present or nonpresent).

Implementation

To approach this problem, we implemented a program in the *R* programming language that accepted minor allele frequencies *p*, relative risks γ, and number of SNPs *n*. This program made it much easier to calculate the values of F/R for comparison, particularly for the *n*-SNP case, which would be nearly impossible by hand calculation.

We treated each SNP's genotype as a binary value, that is, '1' for mutation presence, and '0' for normal. For example, applying the formula from the previous slide to a 2-SNP case, and treating all p_i's and γ_i's as the same values *p* and γ, then as seen before:

- $F/R = (1-P_1)(1-P_2) + P_1\gamma_1(1-P_2) + P_2\gamma_2(1-P_1) + P_1\gamma_1P_2\gamma_2$ = 0 0 + 0 1 + 1 0 + 1 1
  - Which simplifies to: $F/R = (1-P)^2 + P\gamma(1-P) + (1-P)P\gamma + (P\gamma)^2$

To assure that we tested all possible combinations of genotypes for *n* SNPs, we generated an n-digit binary matrix, which was composed of all n-digit binary numbers with values from 0 to 2^n-1 (decimal). Hence, there were 2^n rows, and *n* columns in this matrix. Each row represented one of the terms in the F/R equation, and each column represented a SNP. We then iterated over each row, counting the number of '1' and '0' values. For each '1' value in the row, we multiplied the term by Pγ, and for each '0' value in the row, we multiplied the term by (1-P). Finally, we summed all of the terms, represented by each row, to get the final value of F/R.

Methods

We calculated the F/R (genetic contribution) for several cases:

- Holding the # of SNPs constant, how do varying minor allele frequencies affect F/R?

- Holding the # of SNPs constant, how do varying relative risks for each SNP affect F/R?
- Finally, holding the minor allele frequencies and relative risks constant for all SNPs, how does the number of SNPs affect F/R?

We then compared and contrasted the three scenarios.

Results

In order to get a clearer look at the results of the analysis, I recommend looking at the tables and graphs included in the slide presentation, which can be found at this project's wiki page, which is located at: http://cs124project.wikidot.com/p-med.

Conclusions

We were able to draw several interesting conclusions from the analysis. Firstly, as expected, when holding all else constant, a larger minor allele frequency lead to an increased genetic contribution to disease risk. Likewise, all else held constant, larger relative risks also lead to increased genetic contribution to disease risk. Similarly, larger numbers of SNPs leads to an increased genetic factor of disease risk.

However, the three factors do not scale equivalently. For example, holding p=0.1, increasing the γ from 1 to 10 (a factor of 10) increased F/R by factors of 1.9, 3.61, and 613, for one, two, and ten SNPs, respectively. Yet similarly holding p=0.3 and increasing the number of SNPs from 1 to 10 increased F/R by factors of 2.35, 68.68, 1207.27, and 130,000 for γ=2, 3, 5, and 10, respectively. We can see from the graph that increasing the number of SNPs by a factor of 10 has a greater effect on F/R than increasing γ in most cases .

As another example, we had two SNPs, each with p=0.1 and γ=10, which gave an F/R value of 3.61. A single SNP with the same *p* hoping to achieve the same F/R value would require a relative risk of 25. Going from p=0.1 to p=0.5 (a factor of five) and holding all else equal results in greater F/R gains than going from γ=1 to γ=5 (also a factor a five), the majority of the time. However, this does not apply to cases of low γ values, and as such, this does not hold.

Overall, we were successful in creating a rudimentary program that allows for the entry of minor allele frequency, relative risk, and number of SNPs, and can calculate the genetic contribution to disease risk from these numbers. We were also successful in running several sets of numbers, and getting a good look at how each of these factors ultimately affected the F/R value.

Future Work

In the future, several improvements could be made to the R programming, which would allow us to expand on the study. First of all, scalability was an issue; the implementation of the program in *R* doesn't allow for the creation of large matrices, such as in cases where $n > 20$, due to a memory bound. This might be improved by altering the algorithm that generates the binary matrix. Secondly, the program could be altered to allow for differing values of *p* and *γ*, which would give a more realistic approximation of a disease risk. Perhaps we could allow the input of several vectors of *p* and *γ* values, which would then allow for real-world data input.