Genetic complexity of hypertrophic cardiomyopathy revealed by high throughput sequencing

November 7, 2012

This document provides the mathematical details for the statistical issues that arose in the manuscript entitled *Genetic complexity of hypertrophic cardiomyopathy revealed by high throughput sequencing.*

1 Bringing the data into the R software

We first load the data into R. The numbers below indicate, for each gene, the frequencies of candidate variants (i.e. non-synonymous and rare in the 1,000 genomes dataset) cases and controls.

```
> n1 <- 1287
> n2 <- 180
> genes <- data.frame(gene = c('MYH7', 'TNNT2', 'TNNI3', 'MYBPC3', 'MYL2', 'MYL3', 'ACTC1', 'TPM1'),
+                     f.controls = c(0.0125, 0.00155, 0.00000, 0.02725, 0.0035, 0.00194, 0.00039, 0.001
+                     f.cases = c(0.086, 0.027, 0.0083, 0.05, 0.011, 0.0055, 0.0027, 0.0027))
> genes

    gene f.controls f.cases
1   MYH7    0.01250  0.0860
2  TNNT2    0.00155  0.0270
3  TNNI3    0.00000  0.0083
4 MYBPC3    0.02725  0.0500
5   MYL2    0.00350  0.0110
6   MYL3    0.00194  0.0055
7  ACTC1    0.00039  0.0027
8   TPM1    0.00136  0.0027
```

2 Estimating the proportion of cases explained for each gene

The most difficult question is the following: for a gene of interest, and given the estimated frequencies of candidate variants in controls and cases, how can we estimate the confidence interval for the proportion of HCM cases explained by this class of variants in that gene? Mathematically, we use the following notations: x is the frequency of candidate variants in controls and $x + y$ is the frequency of candidate variants in cases. The parameter y refers to the additional proportion of cases whose disease is caused by mutations in that gene. What we are after is a confidence interval for y.

Here, X is a nuisance parameter and we use a profile likelihood argument to get rid of this parameter. Precisely, we set up a grid of values for y and for each of these we find the value of x that maximizes the lieklihood of the data. The function that we use to obtain the confidence interval for the parameter y is below. $n_1 = 1$, is the number of controls and $n_2 = 180$ is the number of cases.

```
> joint.estimate <- function(n1, n2, x1, x2) {
+   y <- seq(0, 3*x2, x2/100)  ##grid of values for y
+
+   ## Now here is the likelihood function
+   my.fn <- function(x, y) ( -floor(x1*n1)*log(x)
+                             - (n1 - floor(x1*n1))*log(1-x)
+                             - floor(x2*n2)*log(x+y)
+                             - (n2 - floor(x2*n2))*log(1- x - y))
+   prof <- rep(NA, length(y))  ##profile likelihood
+
+   ## Now loop over all possible values of y and maximise wrt x
+   for (i in 1:length(y)) {
+     prof[i] = optim(par = x1,      ##here is the maximisation for x
+           fn = my.fn,
+           gr = NULL,
+           y = y[i],
+           lower = 0.001,
+           upper = 1 - y - 0.001,
+           method = 'Brent')$value
+   }
+   prof <- ifelse (is.nan(prof), -10^6, prof)
+   prof <- -prof
+   prof <- prof - max(prof)
+
+   in.CI <- which(2*prof > -qchisq(df = 1, p = 0.95, lower.tail = TRUE))
+   my.CI <- y[ range( in.CI ) ]
+   return(my.CI)
+ }
```

We can now estimate this confidence interval for each gene in our dataset. Because each control/case has 2 chromosomes, if the frequency of candidate variants is f, then the proportion of cases with at least one causal variant is $f^2 + 2f(1-f)$.

```
> for  (i in 1:nrow(genes)) {
+   prop.controls <- 2*genes$f.controls[i]*(1-genes$f.controls[i]) + genes$f.controls[i]^2
+   prop.cases <- 2*genes$f.cases[i]*(1-genes$f.cases[i]) + genes$f.cases[i]^2
+
+   my.CI <- joint.estimate (n1 = n1, n2 = n2, x1 = prop.controls, x2 = prop.cases)
+   my.CI <- signif(my.CI, 3)
+   genes$CI.prop.HCM.cases.explained[i] <- paste( '[', my.CI[1], '-', my.CI[2], ']', sep = '')
+ }
> genes
```

```
    gene f.controls f.cases CI.prop.HCM.cases.explained
1   MYH7    0.01250  0.0860              [0.0889-0.196]
2  TNNT2    0.00155  0.0270             [0.0218-0.0858]
3  TNNI3    0.00000  0.0083           [0.000992-0.0329]
4 MYBPC3    0.02725  0.0500             [0.00195-0.0897]
5   MYL2    0.00350  0.0110                   [0-0.0365]
6   MYL3    0.00194  0.0055                   [0-0.0213]
7  ACTC1    0.00039  0.0027                   [0-0.0106]
8   TPM1    0.00136  0.0027                   [0-0.0106]
```

3 P-values for differences between cases and controls

This computation is done by statistically comparing the number of candidate variants between cases and controls. We use a Fisher exact test to quantify this difference. Because we work at the allele level the observed number of alleles are $2n_1 * f_1$ in controls and $2n_1 * f_2$ in cases, where f_1 and f_2 denote the frequency of candidate variants in cases and controls.

```
> for  (i in 1:nrow(genes)) {
+   my.mat <- matrix(data = c(2*n1*(1 - genes$f.controls[ i ]),
+                       2*n1*genes$f.controls[i],
+                       2*n2*(1-genes$f.cases[i]),
+                       2*n2*genes$f.cases[i]),
+                     nrow = 2,
+                     ncol = 2,
+                     byrow = TRUE)
+
+   my.mat <- round(my.mat)
+   genes$P.diff[i] <- fisher.test ( my.mat )$p.value
+ }
> genes
```

```
    gene f.controls f.cases CI.prop.HCM.cases.explained      P.diff
1   MYH7    0.01250  0.0860              [0.0889-0.196] 3.855666e-13
2  TNNT2    0.00155  0.0270             [0.0218-0.0858] 4.373726e-07
3  TNNI3    0.00000  0.0083           [0.000992-0.0329] 1.833766e-03
4 MYBPC3    0.02725  0.0500             [0.00195-0.0897] 2.987798e-02
5   MYL2    0.00350  0.0110                   [0-0.0365] 6.473193e-02
6   MYL3    0.00194  0.0055                   [0-0.0213] 2.083091e-01
7  ACTC1    0.00039  0.0027                   [0-0.0106] 2.303803e-01
8   TPM1    0.00136  0.0027                   [0-0.0106] 4.805606e-01
```

```
>
```

4 Point estimates for the probability that a candidate variant is causal

We want to estimate, for each gene, the probability that a candidate variant in a HCM case is causal. This point estimate is straightforward. It is simply:

$$p_{causal} = \frac{\text{Frequency of candidates variants in cases} - \text{Frequency of candidates variants in controls}}{\text{Frequency of candidate variants in cases}}$$

Hence:

```
> genes$prob.causal <- (genes$f.cases - genes$f.controls)/genes$f.cases
> genes
```

```
    gene f.controls f.cases CI.prop.HCM.cases.explained      P.diff
1   MYH7    0.01250  0.0860              [0.0889-0.196] 3.855666e-13
2  TNNT2    0.00155  0.0270             [0.0218-0.0858] 4.373726e-07
3  TNNI3    0.00000  0.0083           [0.000992-0.0329] 1.833766e-03
4 MYBPC3    0.02725  0.0500             [0.00195-0.0897] 2.987798e-02
```

```
5   MYL2    0.00350  0.0110                [0-0.0365] 6.473193e-02
6   MYL3    0.00194  0.0055                [0-0.0213] 2.083091e-01
7  ACTC1    0.00039  0.0027                [0-0.0106] 2.303803e-01
8   TPM1    0.00136  0.0027                [0-0.0106] 4.805606e-01
  prob.causal
1   0.8546512
2   0.9425926
3   1.0000000
4   0.4550000
5   0.6818182
6   0.6472727
7   0.8555556
8   0.4962963
```