Elinor Velasquez, elinor.velasquez@Berkeley.edu

Computing the PRS

A perusal of the literature (PubMed) showed two main methods for computing the polygenic risk score:

1. PRS = Σ G(i)β(i), G(i) is associated with the risk allele or minor allele, and β(i) is a coefficient (could be learned if not given). One issue is that the risk allele is not always equal to the minor allele.
2. Prune and Threshold: A variety of software is available for this type of computation. Vilhjálmsson et al. [2], construct an algorithm based upon linkage disequilibrium considerations: The resulting software, LD_{pred}, is contained in the Basic Tutorial for Polygenic Risk Score Analyses [3], and can also be found in the GitHub repository.

I chose the first method (Dudbridge, 2013), with $G(i) = (X(i) - 2f(i))/(2f(i)(1 - f(i)))^{1/2}$ where X(i) is the number of minor alleles per patient and f(i) is the minor allele frequency at SNP i [1].

If the risk allele is not equal to the minor allele in an instance, the risk allele should be chosen in place of the minor allele, in order to increase the risk of thyroid cancer. This choice will be implemented in later versions, pending discussion.

References

1. Dudbridge F (2013) Power and Predictive Accuracy of Polygenic Risk Scores. PLoS Genet 9(3): e1003348.
2. Vilhjálmsson B et al. (2015). Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. Am J Hum Genet *97*, 576–592.
3. https://choishingwan.github.io/PRS-Tutorial/plink/

Additional Questions

1. What type(s) of analysis would you consider for evaluating the relationship between the thyroid cancer PRS and thyroid cancer status? Explain.

One linear model, logistic regression, has already been used to examine the association between the thyroid cancer PRS and thyroid cancer status [1]. Nonlinear classifiers may also be considered to model PRS and thyroid cancer status: The goal is to find a model with the highest accuracy score or least error under supervised learning. A support vector machine classifier with a non-trivial kernel, the CatBoost classifier, the Long

Short-Term Memory (recurrent neural network) classifier are some of the many models (in this case, univariate) that may be applied to model the association between PRS and thyroid cancer status.

2. What type(s) of analysis would you consider for evaluating the relationship between the thyroid cancer PRS and a continuous cognition score? Explain.

The first question dealt with classifiers because thyroid cancer status is assumed to be a discrete measure. A continuous cognition score is expected to be a continuous measure. A variety of linear and nonlinear regressor models may be tried: Linear regression, support vector machine regressors, the CatBoost regressor, the LSTM (recurrent neural network) regressor are among the many models available for supervised learning.

References

1. Liyanarachchi S. et al. (2020). Assessing thyroid cancer risk using polygenic risk scores. Proc Natl Acad Sci U S A. 117(11): 5997–6002.

R Code

```
library('readxl')
data_dos <- read_excel("./Desktop/R/MockDosageData.xlsx", sheet = 1)
data_var <- read_excel("./Desktop/R/MockVariantData.xlsx", sheet = 1)

cols = c('rs104', 'rs117', 'rs118', 'rs142', 'rs157', 'rs169', 'rs171', 'rs189', 'rs193', 'rs199')

df_dos<-data_dos[,cols] # print(df_dos) #patients=rows, SNPs = cols, shape: 100 x 10

df_var1 <- as.data.frame(data_var)
row.names(df_var1) <- df_var1$SNP
df_var<-df_var1[cols,] # information file restricted to subclass of SNPs

X <- as.data.frame(df_dos)
PRS = 0
score = array(0)
G <- as.data.frame(0,)

# Write output to the file 'output.csv'
sink('./Desktop/R/output.csv')

# iterate over patients and subclass of SNPs
for (j in 1:100) { # 100 patients
```

```
        for (i in 1:10) { # 10 informative SNPs

                if (df_var[i,'RA'] == df_var[i,'Al1']) {
                        # print('risk allele is allele 1')
                        X[j,i] = df_dos[j,i] # rows=patients, cols=cols(SNPs)

                        #print(df_var[i,'MAF']) # minor allele frequency for SNP i
                        # G = genetic markers (Dudbridge 2013):
                        # Dudbridge F (2013) Power and Predictive Accuracy of Polygenic Risk
Scores. PLoS  Genet 9(3): e1003348.
                        G[j,i] = (X[j,i] -
2*(df_var[i,'MAF']))/sqrt((2*df_var[i,'MAF'])*(1-df_var[i,'MAF']))

                        prs[i] = G[j,i]*(df_var[i,'Beta'])
                        PRS = PRS + prs[i]
                }
                else if (df_var[i,'RA'] == df_var[i,'Al2']) {
                        # print('risk allele is allele 2')
                        X[j,i] = 2 - df_dos[j, i] # Assume bi-allelic conditions

                        G[j,i] = (X[j,i] -
2*(df_var[i,'MAF']))/sqrt((2*df_var[i,'MAF'])*(1-df_var[i,'MAF']))

                        prs[i] = G[j,i]*(df_var[i,'Beta'])
                        PRS = PRS + prs[i]
                }
                else {
                        print('error')
                }
        }
        print("PRS for Patient: ")
        print(j)
        print(PRS)
        PRS = 0
}
sink()
```

Output

```
[1] "PRS for Patient: "
[1] 1
[1] 0.860264
[1] "PRS for Patient: "
[1] 2
[1] 0.7887078
[1] "PRS for Patient: "
```

[1] 3
[1] 0.4151395
[1] "PRS for Patient: "
[1] 4
[1] 0.2730736
[1] "PRS for Patient: "
[1] 5
[1] 0.6161106
[1] "PRS for Patient: "
[1] 6
[1] 0.8848415
[1] "PRS for Patient: "
[1] 7
[1] 0.4373005
[1] "PRS for Patient: "
[1] 8
[1] 0.4405446
[1] "PRS for Patient: "
[1] 9
[1] 0.6456274
[1] "PRS for Patient: "
[1] 10
[1] 0.8367269
[1] "PRS for Patient: "
[1] 11
[1] 0.1835948
[1] "PRS for Patient: "
[1] 12
[1] 0.448432
[1] "PRS for Patient: "
[1] 13
[1] 0.4822674
[1] "PRS for Patient: "
[1] 14
[1] 0.170662
[1] "PRS for Patient: "
[1] 15
[1] 0.3021628
[1] "PRS for Patient: "
[1] 16
[1] 0.5801495
[1] "PRS for Patient: "
[1] 17
[1] 0.3191049
[1] "PRS for Patient: "
[1] 18
[1] 0.2322294
[1] "PRS for Patient: "
[1] 19
[1] 0.1931666
[1] "PRS for Patient: "
[1] 20
[1] 0.71771
[1] "PRS for Patient: "
[1] 21
[1] 0.7426488
[1] "PRS for Patient: "

[1] 22
[1] 0.4618482
[1] "PRS for Patient: "
[1] 23
[1] 0.3134693
[1] "PRS for Patient: "
[1] 24
[1] 0.3790028
[1] "PRS for Patient: "
[1] 25
[1] 0.3930855
[1] "PRS for Patient: "
[1] 26
[1] 0.5420281
[1] "PRS for Patient: "
[1] 27
[1] 0.3397307
[1] "PRS for Patient: "
[1] 28
[1] 0.383203
[1] "PRS for Patient: "
[1] 29
[1] 0.4707935
[1] "PRS for Patient: "
[1] 30
[1] 0.7744736
[1] "PRS for Patient: "
[1] 31
[1] 0.6867422
[1] "PRS for Patient: "
[1] 32
[1] 0.3343405
[1] "PRS for Patient: "
[1] 33
[1] 0.6585711
[1] "PRS for Patient: "
[1] 34
[1] 0.7635384
[1] "PRS for Patient: "
[1] 35
[1] 0.7676679
[1] "PRS for Patient: "
[1] 36
[1] 0.4030109
[1] "PRS for Patient: "
[1] 37
[1] 0.2246301
[1] "PRS for Patient: "
[1] 38
[1] 0.5615206
[1] "PRS for Patient: "
[1] 39
[1] 0.172109
[1] "PRS for Patient: "
[1] 40
[1] 0.5795908
[1] "PRS for Patient: "

[1] 41
[1] 0.5898801
[1] "PRS for Patient: "
[1] 42
[1] 0.4665639
[1] "PRS for Patient: "
[1] 43
[1] 0.538367
[1] "PRS for Patient: "
[1] 44
[1] 0.5042215
[1] "PRS for Patient: "
[1] 45
[1] 0.5514806
"PRS for Patient: "
[1] 46
[1] 0.5255252
[1] "PRS for Patient: "
[1] 47
[1] 0.06557859
[1] "PRS for Patient: "
[1] 48
[1] 0.4054475
[1] "PRS for Patient: "
[1] 49
[1] 0.5795908
[1] "PRS for Patient: "
[1] 50
[1] 0.3976207
[1] "PRS for Patient: "
[1] 51
[1] 0.6334751
[1] "PRS for Patient: "
[1] 52
[1] 0.5345538
[1] "PRS for Patient: "
[1] 53
[1] 0.5126698
[1] "PRS for Patient: "
[1] 54
[1] 0.6523373
[1] "PRS for Patient: "
[1] 55
[1] 0.1509253
[1] "PRS for Patient: "
[1] 56
[1] 0.7268718
[1] "PRS for Patient: "
[1] 57
[1] 0.3095043
[1] "PRS for Patient: "
[1] 58
[1] 0.5565378
[1] "PRS for Patient: "
[1] 59
[1] 0.3676828
[1] "PRS for Patient: "

[1] 60
[1] 0.7357809
"PRS for Patient: "
[1] 61
[1] 0.2467905
[1] "PRS for Patient: "
[1] 62
[1] 0.7902157
[1] "PRS for Patient: "
[1] 63
[1] 0.2914167
[1] "PRS for Patient: "
[1] 64
[1] 0.2877758
[1] "PRS for Patient: "
[1] 65
[1] 0.3079209
[1] "PRS for Patient: "
[1] 66
[1] 0.4691969
[1] "PRS for Patient: "
[1] 67
[1] 0.4689362
[1] "PRS for Patient: "
[1] 68
[1] 0.1241064
[1] "PRS for Patient: "
[1] 69
[1] 0.2065634
[1] "PRS for Patient: "
[1] 70
[1] 0.2647228
[1] "PRS for Patient: "
[1] 71
[1] 0.6279442
[1] "PRS for Patient: "
[1] 72
[1] 0.3438857
[1] "PRS for Patient: "
[1] 73
[1] 0.2933552
[1] "PRS for Patient: "
[1] 74
[1] 0.4927633
[1] "PRS for Patient: "
[1] 75
[1] 0.7190901
"PRS for Patient: "
[1] 76
[1] 0.2272132
[1] "PRS for Patient: "
[1] 77
[1] 0.4000925
[1] "PRS for Patient: "
[1] 78
[1] 0.5612477
[1] "PRS for Patient: "

[1] 79
[1] 0.5239917
[1] "PRS for Patient: "
[1] 80
[1] 0.7500051
[1] "PRS for Patient: "
[1] 81
[1] 0.3621687
[1] "PRS for Patient: "
[1] 82
[1] 0.5364474
[1] "PRS for Patient: "
[1] 83
[1] 0.6424602
[1] "PRS for Patient: "
[1] 84
[1] 0.636048
[1] "PRS for Patient: "
[1] 85
[1] 0.8608443
[1] "PRS for Patient: "
[1] 86
[1] 0.3812706
[1] "PRS for Patient: "
[1] 87
[1] 0.5146083
[1] "PRS for Patient: "
[1] 88
[1] -0.176743
[1] "PRS for Patient: "
[1] 89
[1] 0.780357
[1] "PRS for Patient: "
[1] 90
[1] 0.450124
[1] "PRS for Patient: "
[1] 91
[1] 0.6710925
[1] "PRS for Patient: "
[1] 92
[1] 0.6164147
[1] "PRS for Patient: "
[1] 93
[1] 0.08670022
[1] "PRS for Patient: "
[1] 94
[1] 0.221799
[1] "PRS for Patient: "
[1] 95
[1] 0.3127204
[1] "PRS for Patient: "
[1] 96
[1] 0.6654385
[1] "PRS for Patient: "
[1] 97
[1] 0.3021628
[1] "PRS for Patient: "

[1] 98
[1] 0.09421095
[1] "PRS for Patient: "
[1] 99
[1] 0.8183622
[1] "PRS for Patient: "
[1] 100
[1] 0.5648885