Low Polygenic Risk Scores Among Cancer Patients Inform the Presence of Hereditary Cancer Syndromes

Iris Gupta, Dr. Manish Gala
Harvard Medical School and Mass. Gen. Hospital

Abstract

Colorectal cancer (CRC) is the third most deadly cancer in the world. While Genome-wide (GWAS) have been instrumental in identifying common pathogenic variants that increase risk for colorectal cancer and Polygenic Risk Scores (PRS) have been used to quantify a patient's risk for developing cancer, it has not yet investigated how PRS should be viewed in cases where patients deemed lower-risk develop cancer. The focus of this work was to identify the relative prevalence of pathogenic rare variants in CRC predisposition, which are currently excluded from conventional GWAS and PRS, among quartiles of PRS. I analyzed the PRS of 563, in conjunction with whole exome data for colorectal cancer patients from The Cancer Genome Atlas (TCGA), controlled for relevant epidemiological risk factors, and performed a multivariable logistic regression analysis to determine if pathogenic rare variants in CRC predisposing genes are statistically linked with patients whose PRS are in the lowest quartile. Results showed patients in the lowest quartile of PRS had the greatest effect size in predicting the presence of pathogenic rare variants within the cohort, although the results are not conclusive. Understanding why PRS models fail in the prediction of CRC will pave the way to identify actionable, high-risk variants for cancer predisposition that can influence screening strategies for patients and their families.

Background

- Genome wide association studies (GWAS) are used to identify associations between gene loci and phenotypic traits
- Common disease-common variant (CDCV) hypothesis posits that common genetic variants compromise much of the heritability for common diseases
- Polygenic risk scores (PRS) are used to estimate the combined effect of multiple variants on risk for developing a common trait identified by GWAS
- A PRS in the highest quartile will indicate a heightened genetic risk for the trait compared to a score in a lower quartile
- Colorectal cancer (CRC) is a complex disease, meaning it is caused by the interaction of multiple genes and environmental factors
- GWAS for CRC identified 95 pathogenic common variants
- Accurate CRC risk prediction models are critical for identifying individuals at low and high risk of developing CRC, as high-risk groups can be targeted for screening and chemoprevention strategies
- Little has been discussed how PRS should be viewed in the context of when individuals with low PRS develop CRC
- My goal was to understand why PRS fail in the prediction of CRC

Results

Polygenic Risk Scores Distribution

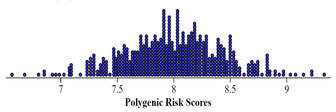


Figure 1. A visual distribution of polygenic risk scores for 563 colorectal cancer patients.

Normality of Polygenic Risk Scores

```
        Shapiro-Wilk normality test

data:  my_data$Scores
W = 0.99812, p-value = 0.8012
```

Figure 2. The R-Studio output of a Shapiro-Wilk Normality Test on polygenic risk scores.

According to Figure 2, the p-value of a Shapiro-Wild normality test was p = 0.8012, indicating an approximately normal distribution.

Normality of Polygenic Risk Scores

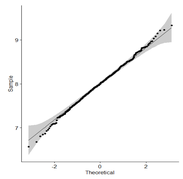
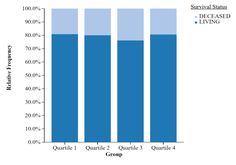


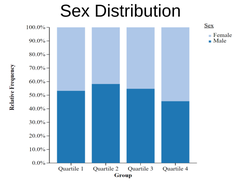
Figure 3. The R-Studio output of a Q-Q plot using polygenic risk scores.

Figure 3 displays another method of analyzing normally using R. The Q-Q plot allowed for visualization of how far data points stray from normality, which is indicated by the diagonal line.

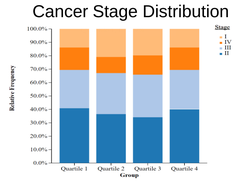
Survival Status Distribution



Sex Distribution



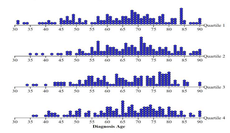
Cancer Stage Distribution



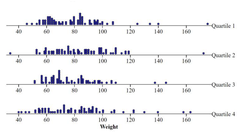
Ancestry Distribution

Figures 4, 5, 6, 7. The distribution of categorical epidemiological variables amongst four PRS quartiles using a Chi-Square test of homogeneity.

Diagnosis Age Distribution



Weight Distribution



Figures 4, 5, 6, and 7 show the results of a Chi-Square analysis for the variables Survival Status, Sex, Cancer Stage, and Ancestry, respectively. The p-values calculated were 0.753, 0.2, 0.725, and 0.009, respectively, indicating that only Ancestry showed a significantly varying distribution amongst quartiles.

Figures 8 and 9 show the results of a One-Way ANOVA Test for the variables Diagnosis Age and Weight, respectively, indicating that both variables showed an approximately equal distribution amongst quartiles.

Figures 8, 9. The distribution of quantitative epidemiological variables amongst four PRS quartiles using a One-Way ANOVA Test.

IGV Snapshot Example

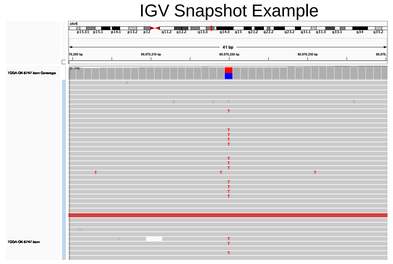


Figure 10. An example of an IGV snapshot used for pathogenic rare variant identification and analysis.

Figure 10 shows an IGV snapshot that was used to identify rare variants. This particular variant was deemed "useful" as it has over 20 readings (rows) and the "T" substitution mutation at the shown genomic location (red column) is considered pathogenic.

Identified Pathogenic Rare Variants in Colorectal Cancer Patients

Gene	SNP Location	rsID
CHEK2	28695868	rs555607708
CHEK2	28695868	rs555607708
GALNT12	98831882	rs746110126
MLH1	37012098	rs63751615
MSH2	47476399	rs63749932
MSH2	47429885	rs760228651
MSH3	80670220	rs371356175
MSH6	47798724	rs587781691
MSH6	47799588	rs863224829
MUTYH	45331566	rs36053993
MUTYH	45332952	rs771184451
MUTYH	45332056	rs776362892
MUTYH	45331566	rs36053993
NTHL1	2046238	rs150766139
NTHL1	2040004	rs148347092
NTHL1	2046238	rs150766139

Table 1. The 16 identified pathogenic rare variants.

Table 1 shows the 16 pathogenic rare variants that were identified through IGV. Each variant is categorized by the gene it is located on, its location in said gene, and its rsID. Repeated variants indicate that more than one patient expressed that variant in their genome.

Coefficients

Term	Coef	SE Coef	Z-Value	P-Value	VIF
Constant	-12	134	-0.09	0.927	
Diagnosis Age	-0.0526	0.0216	-2.43	0.015	1.15
Sex					
Male	-0.482	0.591	-0.82	0.414	1.06
Stage					
II	0.457	0.844	0.54	0.588	2.16
III	-0.042	0.902	-0.05	0.963	2.11
IV	-0.92	1.26	-0.73	0.466	1.42
Ancestry					
African-European	11	134	0.08	0.933	14740.85
Asian	-0	268	-0.00	0.999	1.33
European	11	134	0.09	0.931	14741.14
Quartile Reversed					
2	0.544	0.933	0.58	0.560	1.91
3	0.31	1.02	0.30	0.765	1.72
4	1.364	0.851	1.60	0.109	2.19

Table 2. The coefficients and p-values of each variable included in the binary regression model in predicting the presence of a pathogenic rare variant.

Table 2 shows the MiniTab output of the binary regression model performed on the 563 colorectal cancer patients. The binary outcome of the regression was "Pathogenic Rare Variant" and the inputs of the regression were Diagnosis Age, Sex, Stage, Quartile, and Ancestry. Quartile 4 was used as the reference category for the variable "Quartile". The Quartile output in Table 3 is reversed in order to ensure Quartile 4 was used as the reference. For example, Quartile 1 = "Quartile 4" in the output, Quartile 2 = "Quartile 3" in the output. The p-values calculated from the binary logistic regression for Quartile 1, Quartile 2, Quartile 3, and Quartile 4 were 0.109, 0.765, 0.560, and 1.000, respectively. Coefficients for each quartile were further analyzed to determine effect size. The regression equation coefficients for Quartile 1, Quartile 2, and Quartile 3 were 1.364, 0.31, and 0.544, respectively.

Methods

- TCGA data for 563 CRC patients were genotyped and imputed from the Genomic Data Commons Data Portal
- PRS were calculated for all patients in the cohort and the normality of their distribution was determined
- PRS were split into four quartiles in ascending order
- The distributions of variables: Survival status, ancestry, cancer stage, sex, weight, and diagnosis age were analyzed
- Rare variants were identified amongst patients in the cohort in IGV using the following criteria: IGV Readings > 20 & variants are deemed pathogenic in pre-existing literature
- A binary logistic regression analysis was performed using the input variables: Sex, cancer stage, ancestry, PRS quartile, and the binary output variable "Presence of a Pathogenic Rare Variant"

Conclusions

- The distribution of PRS was approximately normal
- The distributions for the variables: Survival status, sex, stage, diagnosis age, and weight were approximately equal amongst the 4 quartiles of PRS
- The variable Ancestry had a significantly different distribution amongst the 4 quartiles
- 16 pathogenic rare variants were identified amongst the 563 CRC patients
- Although pathogenic rare variants were not significantly linked with low-quartile patients, their effect size is the greatest for this cohort
- Overall, there is enough evidence to justify an alteration of the CDCV and conclude that low PRS inform the presence of pathogenic rare variants

Future Research

- Extend my investigation into pathogenic rare variants to other complex diseases such as breast cancer or Alzheimer's
- Perform my analysis with a more comprehensive gene list
- Account for patients having multiple pathogenic rare variants in their genome as opposed to using a binary variable
- Investigate a broader range of epidemiological variables and incorporate them into my logistic analysis