12/13 update

Linda Gai

12/13/2018

R Markdown

Cleaning vcf

The raw vcf `/dcl01/beaty/data/gmkf/euro/vcfs/filtered/8q24.recode.vcf` was cleaned by:

1. Removing all tri-alellic SNPs
2. Individuals with Mendelian errors
3. Phased using BEAGLE 4.0 (i.e., half calls were not removed).
4. Monomorphic SNPs were removed.

TODO: filter to MAF cut-off to 0.05 or 0.01 (is MAF possibly captured in functional annotation information)?

Filtering

The ultimate goal was to run all 3 tests on the same datasets to compare results. Since rvTDT uses allele frequencies in the population to weight SNPs in the case-parent trios, I first filtered to all SNPs that had allele frequences in Europeans available from 1000Genomes from Aug 2015 (in the Annovar Report). I then created datasets with 3 different types of filters.

1. Filtered by functional annotation information (1048 SNPs)

I filtered to SNPs with scores (in the ANNOVAR report) of CADD > 10, GWAVA > 0.4, EIGEN >4, using recommended cut-offs from literature, or, in the case of EIGEN, a cut-off used in at least one paper. I used these scores because the other annotation scores (SIFT, PolyPhen) were completely missing from the ANNOVAR report. Unfortunately, there was high degree of missingness in each of the scores, so this approach may have excluded some posibly causal SNPs if no annotation score was available for them.

The positions that have been kept have been labeled red in the gTDT plot below.

2. Filtered to region with peak TDT signal (679 SNPs)

The positions that have been kept have been labeled red in the gTDT plot below. Score ranges for .

3. Both 1 and 2 (119 SNPs)

The positions that have been kept have been labeled red in the gTDT plot below.

The results for the 3 tests (rvTDT, RV-TDT, and ScanTrio are given below):

Results

1. Filtered by functional annotation information (CADD > 10, GWAVA > 0.4, EIGEN >4, using literature) using cut-offs (1048 SNPs)

rvTDT

```
summary(cars)
```

```
##      speed           dist
##  Min.   : 4.0   Min.   :  2.00
##  1st Qu.:12.0   1st Qu.: 26.00
##  Median :15.0   Median : 36.00
##  Mean   :15.4   Mean   : 42.98
##  3rd Qu.:19.0   3rd Qu.: 56.00
##  Max.   :25.0   Max.   :120.00
```

RV-TDT

```
summary(cars)
```

```
##      speed           dist
##  Min.   : 4.0   Min.   :  2.00
##  1st Qu.:12.0   1st Qu.: 26.00
##  Median :15.0   Median : 36.00
##  Mean   :15.4   Mean   : 42.98
##  3rd Qu.:19.0   3rd Qu.: 56.00
##  Max.   :25.0   Max.   :120.00
```

ScanTrio

```
summary(cars)
```

```
##      speed           dist
##  Min.   : 4.0   Min.   :  2.00
##  1st Qu.:12.0   1st Qu.: 26.00
##  Median :15.0   Median : 36.00
##  Mean   :15.4   Mean   : 42.98
##  3rd Qu.:19.0   3rd Qu.: 56.00
##  Max.   :25.0   Max.   :120.00
```

2. Filtered to peak TDT signal (679 SNPs)

rvTDT

```
summary(cars)
```

```
##      speed           dist
##  Min.   : 4.0   Min.   :  2.00
##  1st Qu.:12.0   1st Qu.: 26.00
##  Median :15.0   Median : 36.00
##  Mean   :15.4   Mean   : 42.98
##  3rd Qu.:19.0   3rd Qu.: 56.00
##  Max.   :25.0   Max.   :120.00
```

RV-TDT

```
summary(cars)
```

```
##      speed           dist
##  Min.   : 4.0   Min.   :  2.00
##  1st Qu.:12.0   1st Qu.: 26.00
##  Median :15.0   Median : 36.00
##  Mean   :15.4   Mean   : 42.98
##  3rd Qu.:19.0   3rd Qu.: 56.00
##  Max.   :25.0   Max.   :120.00
```

ScanTrio

```
summary(cars)
```

```
##      speed           dist
##  Min.   : 4.0   Min.   :  2.00
##  1st Qu.:12.0   1st Qu.: 26.00
##  Median :15.0   Median : 36.00
##  Mean   :15.4   Mean   : 42.98
##  3rd Qu.:19.0   3rd Qu.: 56.00
##  Max.   :25.0   Max.   :120.00
```

Interpretation:

3. Both 1 and 2 (119 SNPs)

rvTDT

```
summary(cars)
```

```
##      speed           dist
##  Min.   : 4.0   Min.   :  2.00
##  1st Qu.:12.0   1st Qu.: 26.00
##  Median :15.0   Median : 36.00
##  Mean   :15.4   Mean   : 42.98
##  3rd Qu.:19.0   3rd Qu.: 56.00
##  Max.   :25.0   Max.   :120.00
```

RV-TDT

```
summary(cars)
```

```
##      speed           dist
##  Min.   : 4.0   Min.   :  2.00
##  1st Qu.:12.0   1st Qu.: 26.00
##  Median :15.0   Median : 36.00
##  Mean   :15.4   Mean   : 42.98
##  3rd Qu.:19.0   3rd Qu.: 56.00
##  Max.   :25.0   Max.   :120.00
```

ScanTrio

```r
summary(cars)
```

```
##      speed           dist       
##  Min.   : 4.0   Min.   :  2.00  
##  1st Qu.:12.0   1st Qu.: 26.00  
##  Median :15.0   Median : 36.00  
##  Mean   :15.4   Mean   : 42.98  
##  3rd Qu.:19.0   3rd Qu.: 56.00  
##  Max.   :25.0   Max.   :120.00
```

Interpretation:

Previous results

I also tried using windows and filtering on RV-TDT over the summer. I first filtered to all positions with a CADD score >10, the recommended minimum cut-off, and then used relatively large windows of 76 or 100 SNPs, with either no overlap or a 10 SNP overlap.

Overall, it did not seem like there

Results are here:

Whole 8q24 region, no overlap, 100 SNPs per window

Whole 8q24 region, 10 overlap, 100 SNPs per window

```
## #gene    CMC-Analytical  BRV-Haplo   CMC-Haplo   VT-BRV-Haplo    VT-CMC-Haplo    WSS-Haplo    8q24
## 
## #gene    CMC-Analytical  BRV-Haplo   CMC-Haplo   VT-BRV-Haplo    VT-CMC-Haplo    WSS-Haplo    8q24
## 
## #gene    CMC-Analytical  BRV-Haplo   CMC-Haplo   VT-BRV-Haplo    VT-CMC-Haplo    WSS-Haplo    8q24
## 
## #gene    CMC-Analytical  BRV-Haplo   CMC-Haplo   VT-BRV-Haplo    VT-CMC-Haplo    WSS-Haplo    8q24
## 
## #gene    CMC-Analytical  BRV-Haplo   CMC-Haplo   VT-BRV-Haplo    VT-CMC-Haplo    WSS-Haplo    8q24
## 
## #gene    CMC-Analytical  BRV-Haplo   CMC-Haplo   VT-BRV-Haplo    VT-CMC-Haplo    WSS-Haplo    8q24
## 
## #gene    CMC-Analytical  BRV-Haplo   CMC-Haplo   VT-BRV-Haplo    VT-CMC-Haplo    WSS-Haplo    8q24
## 
## #gene    CMC-Analytical  BRV-Haplo   CMC-Haplo   VT-BRV-Haplo    VT-CMC-Haplo    WSS-Haplo    8q24
## 
## #gene    CMC-Analytical  BRV-Haplo   CMC-Haplo   VT-BRV-Haplo    VT-CMC-Haplo    WSS-Haplo    8q24
```

TDT peak, no overlap, 100 SNPs

TDT peak, no overlap, 76 SNPs