Dimension Reduction using Local Principal Components (DRLPC)

Fatemeh Yavartanoo

December, 2024

Dimension Reduction using Local Principal Components (DRLPC) is a method designed for gene-level association analysis. It clusters highly correlated SNPs and replaces each cluster with a local principal component, effectively reducing dimensionality. DRLPC also addresses multi-collinearity by iteratively removing variables with high Variance Inflation Factor (VIF), ensuring robust and interpretable analysis of complex genetic data. In this guide, we illustrate how to use DRLPC, including the necessary input files and threshold parameters. The example dataset is based on the 6,219 biallelic SNPs (MAF>0.05) genotyped in chr22:18600803-21496954 in 503 individuals from whole genome sequnceed 1000 Genome Projects European ancestry. In the second step of DRLPC, highly correlated SNPs are clustered based on their LD block using the BigLD algorithm (http://pubmed.ncbi.nlm.nih.gov/29028986/) implemented in R packgae gpart (http://academic.oup.com/bioinformatics/article/35/21/4419/5487391).

Main inputs:

Input 1: SNPinfo This dataframe must include at least chr, rsID and position (bp).

```
> head(SNPinfo)
  chrN       rsID       bp
1   22   rs362108 18600803
2   22   rs361540 18601415
3   22   rs362207 18603088
4   22   rs458471 18605745
5   22 rs62229475 18605770
6   22 rs62229476 18605772
```

Input 2: genoout This dataframe should contain the genotypes (columns) for the individuals (rows). The column names in genoout must match the rsID column in SNPinfo and follow the same order.

```
> head(genoout[,1:6])
     rs362108 rs361540 rs362207 rs458471 rs62229475 rs62229476
[1,]        0        1        0        0          0          0
[2,]        0        1        1        1          0          0
[3,]        1        2        2        2          2          2
[4,]        0        0        0        0          0          0
[5,]        1        2        1        1          1          1
[6,]        0        1        0        0          0          0
```

Partitioning Chromosome 22 by Genes

We partitioned chromosome 22 by genes, using a geneSNPinfo which is a dataframe including gene names and SNP positions (with rsIDs) grouped by their associated genes.
The geneSNPinfo filewas created using the Ensembl BioMart tool by specifying the chromosome, gene name, gene start_bp, and gene end_bp and combining this information with the SNPinfo file that includes SNP positions. The resulting dataframe lists the gene, its genomic positions, and its associated SNPs. In this example, the geneSNPinfo file contains 84 unique genes, with the first gene, "PEX26", including 32 SNPs, the second gene, "TUBA6", including 27 SNPs, and others. This structure facilitates analyses involving genes and their respective SNPs.

```
> geneSNPinfo[30:40,]
   chrN        rsID       bp  gene
30   22 rs148043566 18610616 PEX26
31   22    rs361557 18611223 PEX26
32   22   rs8140197 18613045 PEX26
33   22   rs2540620 18614874 TUBA8
34   22  rs34944413 18616145 TUBA8
35   22   rs5746517 18616930 TUBA8
36   22   rs5993004 18617205 TUBA8
37   22  rs12168977 18618741 TUBA8
38   22   rs9617666 18618851 TUBA8
39   22   rs9618207 18619250 TUBA8
40   22   rs9618208 18619540 TUBA8
```

The geneSNPinfo file is designed to ensure that each gene, along with all SNPs within its boundaries, is analyzed individually in DRLPC. This approach enables dimension reduction to be applied on a gene-by-gene basis, facilitating analysis on biologically meaningful units while simultaneously reducing dimensionality and addressing multicollinearity within each gene.

The DRLPC algorithm requires the following thresholds:

1. **CLQcut**: A threshold for pairwise r^2 values used to cluster highly correlated SNPs through a clique-based graph partitioning method.
2. **VIFcut**: A threshold for the Variance Inflation Factor (VIF) to identify and remove variables with high linear dependencies, addressing multi-collinearity.
3. **PCcut**: A threshold to select Principal Components (PCs) that capture the variability of the removed variables, ensuring the retention of essential genetic information.
4. **Klim**: A partition limit applied during the alias removal step (step 1) to handle subsets of SNPs when their count exceeds the sample size.

These thresholds ensure the effective reduction of dimensionality and multi-collinearity while preserving interpretability and robust analysis.

For this example, we used the following thresholds:

- **CLQcut** = 0.5
- **VIFcut** = 20
- **PCcut** = 0.8
- **Klim** = 300

Main output

The DRLPC generates various outputs that can serve different purposes. However, the primary outputs of DRLPC utilized for regression analysis are as follows:

Output 1: vdata vadat is a dataframe that includes the PHENOTYPE and all final variables, such as SNPs, PCs, and RPCs. (PHENOTYPE can be considered as either a quantitative or qualitative variable.) The following results are for the gene "PEX26" with size 8 after applying DRLPC.

```
> head(DRout$vdata)
   PHENOTYPE SNP23      LPC1       LPC2       LPC3       LPC4       LPC5       LPC6       RPC1
1  0.5255444     1 -2.166557 -0.9798884 -0.6722417 -0.5498254 -0.5719288  0.9606074 -0.4201079
2 -1.1007195     0  1.431189 -0.9798884 -0.6722417 -0.5498254 -0.5719288 -0.3236930 -1.0426222
3 -1.3367770     2  8.831882  7.1311216  4.1906288 -0.5498254 -0.5719288  0.7329644  1.2317669
4 -1.5981963     0 -2.166557 -0.9798884 -0.6722417 -0.5498254 -0.5719288 -1.3803504 -0.4201079
5 -0.3440130     2  2.616295 -0.9798884  2.6617990 -0.5498254 -0.5719288  2.0172648  1.3214931
6  0.6633459     1 -2.166557 -0.9798884 -0.6722417 -0.5498254 -0.5719288 -0.3236930 -0.4201079
```

Output 2: LPCinfo This file contains the Local Principal Components (LPCs) generated for each gene (at step 2 and 3 of the DRLPC), along with all the SNPs included in the analysis. The following result provides detailed LPC information for the gene "PEX26".

```
>   DRout$LPCinfo
                                                    LPC1
"SNP3-SNP4-SNP5-SNP9-SNP17-SNP19-SNP20-SNP24-SNP31-SNP32"
                                                    LPC2
               "SNP8-SNP10-SNP14-SNP16-SNP26-SNP27-SNP29"
                                                    LPC3
                                     "SNP1-SNP15-SNP22"
                                                    LPC4
                                    "SNP13-SNP18-SNP21"
                                                    LPC5
                                          "SNP25-SNP28"
                                                    LPC6
                                            "SNP2-SNP7"
```