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 5.924 biallelic SNPs (MAF>0.05) genotyped in chr22: 18,834,830-21,496,954 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: genotype_data This dataframe should contain the genotypes (columns) for the individuals (rows). The column names in genodata must match the rsID column in SNPinfo and follow the same order.

```
> head(genotype_data[,1:6])
     rs571553209 rs199679028 rs62231277 rs371365085 rs361763 rs201915105
[1,]           0           0          0           0       0           1
[2,]           0           0          0           0       0           1
[3,]           1           1          0           1       0           0
[4,]           0           0          0           0       0           2
[5,]           0           0          1           0       1           0
[6,]           0           0          0           0       0           1
```

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 file was 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 78 unique genes, with the first gene, "AC008132.13", including 32 SNPs, the second gene, "DGCR6", including 27 SNPs, and others. This structure facilitates analyses involving genes and their respective SNPs. We excluded genes with only one SNP, resulting in 72 unique genes.

```
> geneSNPinfo[30:40,]
   chrN        rsID       bp        gene
30   22 rs553696933 18846536 AC008132.13
31   22 rs573530458 18846546 AC008132.13
32   22 rs538813622 18846547 AC008132.13
33   22 rs577634485 18846549 AC008132.13
34   22    rs407895 18893631       DGCR6
35   22    rs416659 18894600       DGCR6
36   22    rs417616 18895006       DGCR6
37   22    rs418623 18895227       DGCR6
38   22   rs1974680 18895563       DGCR6
39   22   rs1974681 18895694       DGCR6
40   22    rs424512 18895703       DGCR6
```

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 multi-collinearity 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
- **PCcut** = 0.8
- **VIFcut** = 20
- **Klim** = 300

Main output

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

Output 1: vdata This is the updated version of the original onedata (containing PHENOTYPE and genotype_data) after dimension reduction. The variables include: remaining SNPs, LPCs (some SNPs are removed and replaced with Local Principal Component (LPC), and Re-added Principal Component (RPC) variables (final step).

Note: PHENOTYPE can be either a quantitative or qualitative variable.

The following results are for the gene "AC004471.10", which originally had 15 SNPs and was reduced to 5 variables after applying DRLPC.

```
> head(DRout$vdata)
   PHENOTYPE SNP9       LPC1       LPC2       LPC3       RPC1
1  -8.087291    0 -1.244984  2.1243402 -0.7106144  1.5249412
2   8.571943    0  3.262637  2.1243402 -0.7106144  2.0270680
3 -11.744968    0  6.910313 -0.7843026 -0.7106144  0.3348369
4  -5.612248    0  2.402692 -0.7843026 -0.7106144 -0.1672899
5  -2.544969    0 -2.104929 -0.7843026  2.3893665 -0.6694166
6  15.388735    0 -2.104929 -0.7843026 -0.7106144 -0.6694166
```

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 "AC004471.10".

```
> DRout$LPCinfo
                                      LPC1
"SNP1-SNP3-SNP6-SNP7-SNP8-SNP11-SNP14"
                                      LPC2
                                "SNP5-SNP10"
                                      LPC3
                                "SNP2-SNP13"
```

Output 3: aliasremoved It includes SNPs which removed due to complete dependency at step 1.

```
> DRout$aliasremoved
[1] "SNP4"  "SNP12" "SNP15"
```

Output 4: removed It includes final removed SNPs after whole procedure.

```
> DRout$removed
[1] "SNP4"  "SNP12" "SNP15"
```

Output 5: RPCind Index of PC variables created and added to the data in final step.

```
> DRout$RPCind
[1] 1
```

Output 6: taglist A vector identifying the tags for removed or replaced SNPs among the remaining variables. The removed SNPs are shown as names, and their corresponding tag variables (e.g., LPCs) are provided as values.

```
> DRout$taglist
  SNP1   SNP2   SNP3   SNP4   SNP5   SNP6   SNP7   SNP8  SNP10  SNP11  SNP12  SNP13  SNP14  SNP15
"LPC1" "LPC3" "LPC1" "LPC1" "LPC2" "LPC1" "LPC1" "LPC1" "LPC2" "LPC1" "LPC2" "LPC3" "LPC1" "LPC1"
```

DRLPC Final Results

The **DRLPC_final_results** data frame is an output generated by the DRLPC algorithm to provide users with a concise summary of the dimension reduction process across all genes. It allows users to quickly review key statistics and assess the effectiveness of DRLPC for each gene.

Columns in DRLPC_final_results:

1. **datanum:** Sequential index of the gene being processed.
2. **genename:** Name of the gene.
3. **size:** The initial number of SNPs before applying DRLPC.

4. **numaliasrmv:** The number of SNPs removed due to aliasing (complete linear dependency) in Step 1.
5. **numremgrps:** The number of remaining groups after clustering highly correlated SNPs (Steps 2 and 3).
6. **maxgrsize:** The size of the largest group of SNPs replaced by a LPC.
7. **numPC**: Number of Re-added Principal Components (RPCs) included in the final dataset (final step).
8. **finalvar:** Final number of variables after applying DRLPC (remaining SNPs, LPCs, and RPCs).

Below is an example of the first few rows of the DRLPC_final_results table:

```
> result[1:10,]
   datanum    genename size numaliasrmv numremgrps maxgrsize numPC finalvar
1        1 AC008132.13   33           1          3         6     0       22
2        2       DGCR6   33           8          4        10     0        6
3        3       PRODH   95          26          8        22     0       14
4        4       DGCR5  182          59         12        37     0       19
5        5      CA15P1    7           1          1         4     0        3
6        6       DGCR2  290         227          4        38     0       11
7        7  AC004471.9    3           0          0         0     0        3
8        8 AC004471.10   15           3          3         7     1        5
9        9      DGCR14   22           7          3        10     1        5
10      10        GSC2    4           1          1         3     0        1
```