240614-

Jingxin Wang

2025-06-13

To Dos

1. Apply INT and log transformation to methylation data and apply classical PCA. Contrast these results with classical PCA on the raw data
2. Read the regional PCA paper in details, identify what they do with the score matrix for each gene region, and identify the research question they are trying to answer.
3. Read over the robust PCA and understand the mathematical principles

Important Update

1. The project will be divided into 4 parts, and we are currently at stage 1 with a brief touch in stage 2.

- Preprocess the data
- Filtering out important genes (for each cancer types)
- Build gene networks
- Compare different networks

2. Regional PCA relies on pre-defined segmentation of genes, but we want to come up with natural groupings of genes. This will be innovation target we try to achieve.

1. What is the goal of the research?

Identify cell-type specific DNA methylation changes that are associated with AD phenotypes.

2. What did they do with the score matrix?

1. Simulation study: to show that rPCA has a greater statistical power than simple averaging in finding the DMRs for every proportion of CpG sites and methlylation differences.
2. Real dataset: the PCA score matrix $Z_r = XV_r$ is calculated for each gene region, and ech column of Z_r is treated as a rPC. Next, a `lmFit` is performed on each score matrix $Z - r$ to identify DMRs.
3. `lmFit` fits multiple linear models by weighted or generalized least squares.

Test Log Transformation and IVT as Data Preprocessing Step

```r
## Function Definition
source("IVT.R")
source("pcaCombo.R")

## Load Data
library(tidyverse)
```

```
## Warning: package 'ggplot2' was built under R version 4.4.3
```

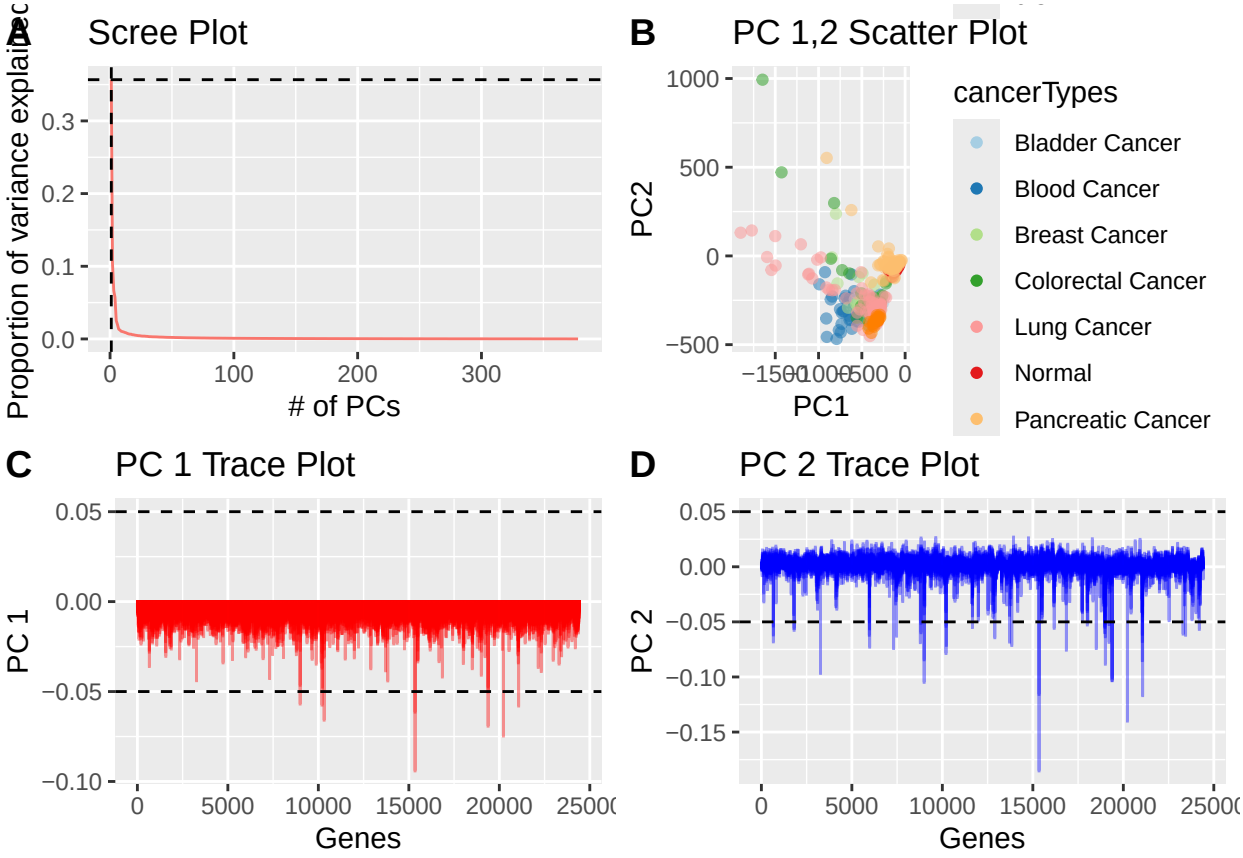
```
## Warning: package 'purrr' was built under R version 4.4.3
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.2     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.4
## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

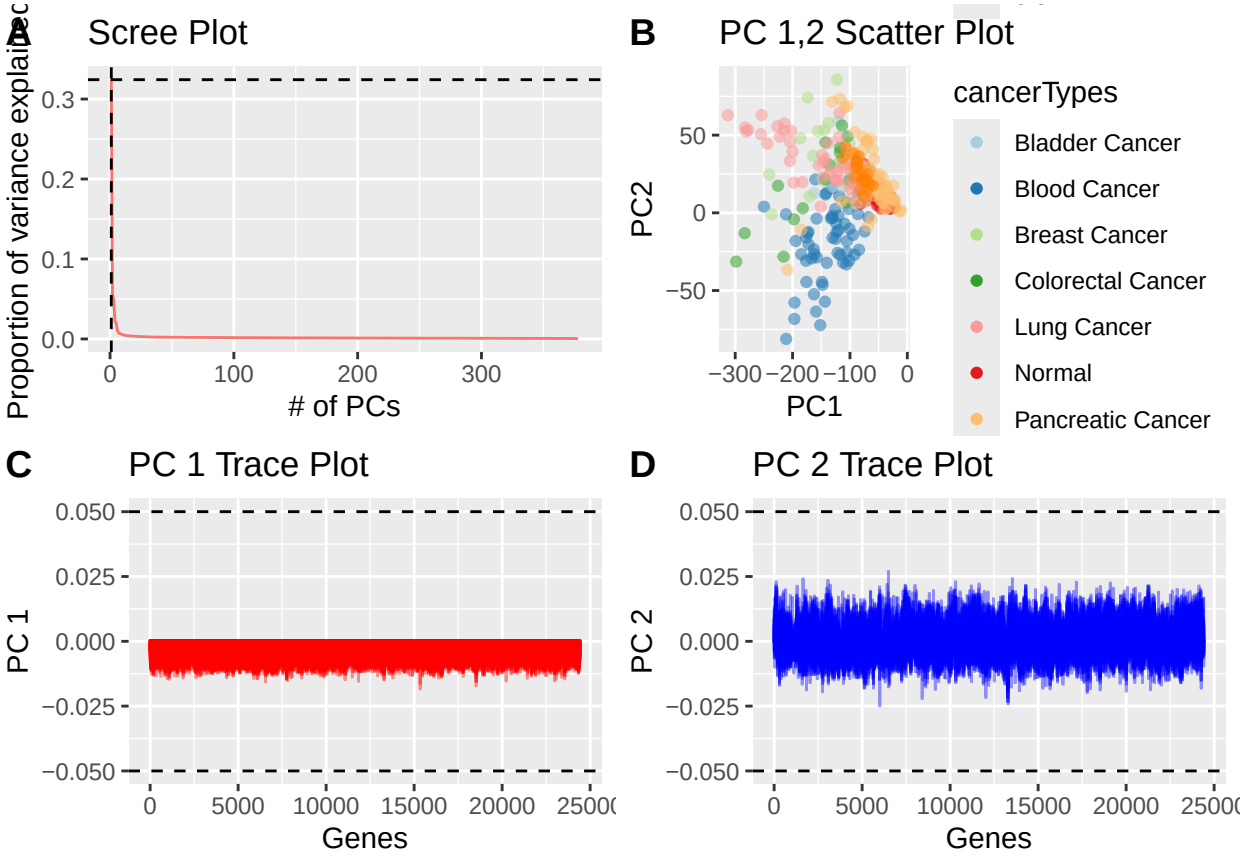
```r
X.T <-  readRDS(file = "../Data/Common_pan_cancer_hyper_bins_adjusted_and_normalized_cnt_in_SE.RDS")
dim(X.T)
```

```
## [1] 24418   378
```

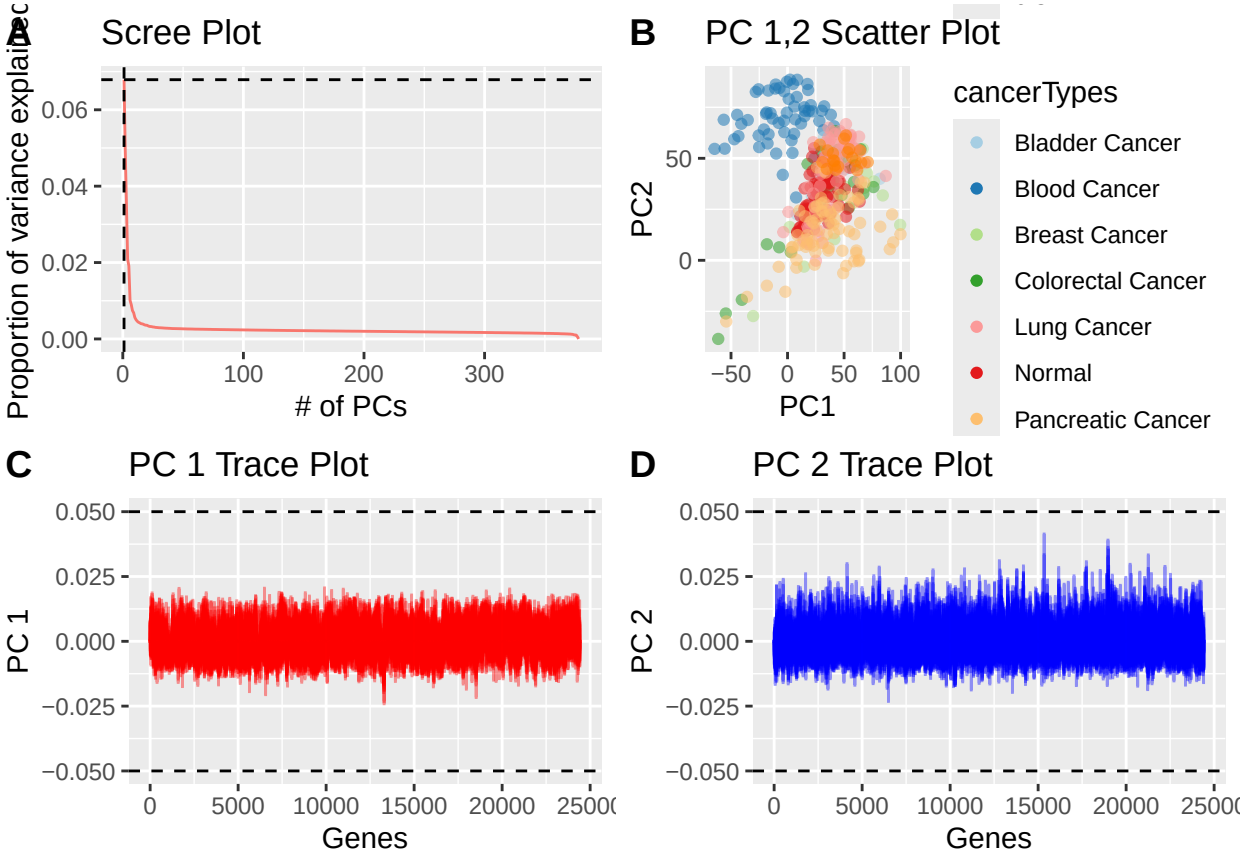
```r
sampleInfo <- read.csv("../Data/Common_pan_cancer_hyper_bins_adjusted_cnt_in_SE_samples_info.csv")
cancerTypes <- sampleInfo$cancer_type[sampleInfo$sample_id %in% colnames(X.T)]

## Raw data without preprocessing
PC_raw = pcaCombo(t(X.T))
```




```
## Log Transformation
logX.T <- log(X.T + 1)
PC_log = pcaCombo(t(logX.T))
```




```
## IVT transformation
k = 0.1
ivtX.T <- IVT(X.T, k)
PC_ivt = pcaCombo(t(ivtX.T))
```



Regional PCA

Robust PCA