BioInformaticsHW3

Divya Boggavarapu(1002086719)

2022-11-07

Collaboration Statement

I have collaborated with Vijitha Kotapati(1001860730) and Tulasi Sridevi Navu-
luru(1002010740), Vandhana(1001876764) to do the programming questions in
1,2,3,4,5,6.

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring
HTML, PDF, and MS Word documents. For more details on using R Markdown see http://
rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as
well as the output of any embedded R code chunks within the document.

Including Plots

You can also embed plots, for example:

```
require(parathyroidSE)

## Loading required package: parathyroidSE

## Loading required package: SummarizedExperiment

## Loading required package: MatrixGenerics

## Loading required package: matrixStats

##
## Attaching package: 'MatrixGenerics'

## The following objects are masked from 'package:matrixStats':
##
##     colAlls, colAnyNAs, colAnys, colAvgsPerRowSet, colCollapse,
##     colCounts, colCummaxs, colCummins, colCumprods, colCumsums,
##     colDiffs, colIQRDiffs, colIQRs, colLogSumExps, colMadDiffs,
##     colMads, colMaxs, colMeans2, colMedians, colMins,
colOrderStats,
```

```
##      colProds, colQuantiles, colRanges, colRanks, colSdDiffs,
colSds,
##      colSums2, colTabulates, colVarDiffs, colVars, colWeightedMads,
##      colWeightedMeans, colWeightedMedians, colWeightedSds,
##      colWeightedVars, rowAlls, rowAnyNAs, rowAnys, rowAvgsPerColSet,
##      rowCollapse, rowCounts, rowCummaxs, rowCummins, rowCumprods,
##      rowCumsums, rowDiffs, rowIQRDiffs, rowIQRs, rowLogSumExps,
##      rowMadDiffs, rowMads, rowMaxs, rowMeans2, rowMedians, rowMins,
##      rowOrderStats, rowProds, rowQuantiles, rowRanges, rowRanks,
##      rowSdDiffs, rowSds, rowSums2, rowTabulates, rowVarDiffs,
rowVars,
##      rowWeightedMads, rowWeightedMeans, rowWeightedMedians,
##      rowWeightedSds, rowWeightedVars

## Loading required package: GenomicRanges

## Loading required package: stats4

## Loading required package: BiocGenerics

##
## Attaching package: 'BiocGenerics'

## The following objects are masked from 'package:stats':
##
##      IQR, mad, sd, var, xtabs

## The following objects are masked from 'package:base':
##
##      anyDuplicated, aperm, append, as.data.frame, basename, cbind,
##      colnames, dirname, do.call, duplicated, eval, evalq, Filter,
Find,
##      get, grep, grepl, intersect, is.unsorted, lapply, Map, mapply,
##      match, mget, order, paste, pmax, pmax.int, pmin, pmin.int,
##      Position, rank, rbind, Reduce, rownames, sapply, setdiff, sort,
##      table, tapply, union, unique, unsplit, which.max, which.min

## Loading required package: S4Vectors

##
## Attaching package: 'S4Vectors'

## The following objects are masked from 'package:base':
##
##      expand.grid, I, unname

## Loading required package: IRanges

##
## Attaching package: 'IRanges'
```

```
## The following object is masked from 'package:grDevices':
##
##     windows

## Loading required package: GenomeInfoDb

## Loading required package: Biobase

## Welcome to Bioconductor
##
##     Vignettes contain introductory material; view with
##     'browseVignettes()'. To cite Bioconductor, see
##     'citation("Biobase")', and for packages 'citation("pkgname")'.

##
## Attaching package: 'Biobase'

## The following object is masked from 'package:MatrixGenerics':
##
##     rowMedians

## The following objects are masked from 'package:matrixStats':
##
##     anyMissing, rowMedians
```

```r
data(parathyroidGenesSE)
parathyroidGenesSE
```

```
## class: RangedSummarizedExperiment
## dim: 63193 27
## metadata(1): MIAME
## assays(1): counts
## rownames(63193): ENSG00000000003 ENSG00000000005 ... LRG_98 LRG_99
## rowData names(0):
## colnames: NULL
## colData names(8): run experiment ... study sample
```

Note that the echo = FALSE parameter was added to the code chunk to prevent printing of the R code that generated the plot.

```r
colData(parathyroidGenesSE)[c(1:10), ]
```

```
## DataFrame with 10 rows and 8 columns
##              run experiment  patient treatment     time submission
study
##     <character>   <factor> <factor>  <factor> <factor>   <factor>
<factor>
## 1    SRR479052  SRX140503        1   Control      24h  SRA051611
```

```
SRP012167
## 2    SRR479053   SRX140504         1   Control      48h   SRA051611
SRP012167
## 3    SRR479054   SRX140505         1   DPN          24h   SRA051611
SRP012167
## 4    SRR479055   SRX140506         1   DPN          48h   SRA051611
SRP012167
## 5    SRR479056   SRX140507         1   OHT          24h   SRA051611
SRP012167
## 6    SRR479057   SRX140508         1   OHT          48h   SRA051611
SRP012167
## 7    SRR479058   SRX140509         2   Control      24h   SRA051611
SRP012167
## 8    SRR479059   SRX140510         2   Control      48h   SRA051611
SRP012167
## 9    SRR479060   SRX140511         2   DPN          24h   SRA051611
SRP012167
## 10   SRR479061   SRX140511         2   DPN          24h   SRA051611
SRP012167
##        sample
##      <factor>
## 1   SRS308865
## 2   SRS308866
## 3   SRS308867
## 4   SRS308868
## 5   SRS308869
## 6   SRS308870
## 7   SRS308871
## 8   SRS308872
## 9   SRS308873
## 10  SRS308873
```

```r
colData(parathyroidGenesSE)[c(11:20), ]
```

```
## DataFrame with 10 rows and 8 columns
##             run experiment   patient  treatment      time  submission
study
##     <character>   <factor>  <factor>   <factor>  <factor>    <factor>
<factor>
## 1    SRR479062   SRX140512         2   DPN          48h   SRA051611
SRP012167
## 2    SRR479063   SRX140513         2   OHT          24h   SRA051611
SRP012167
## 3    SRR479064   SRX140513         2   OHT          24h   SRA051611
SRP012167
```

```
## 4     SRR479065    SRX140514        2    OHT          48h   SRA051611
SRP012167
## 5     SRR479066    SRX140515        3    Control      24h   SRA051611
SRP012167
## 6     SRR479067    SRX140516        3    Control      48h   SRA051611
SRP012167
## 7     SRR479068    SRX140517        3    DPN          24h   SRA051611
SRP012167
## 8     SRR479069    SRX140518        3    DPN          48h   SRA051611
SRP012167
## 9     SRR479070    SRX140519        3    OHT          24h   SRA051611
SRP012167
## 10    SRR479071    SRX140520        3    OHT          48h   SRA051611
SRP012167
##        sample
##      <factor>
## 1   SRS308874
## 2   SRS308875
## 3   SRS308875
## 4   SRS308876
## 5   SRS308877
## 6   SRS308878
## 7   SRS308879
## 8   SRS308880
## 9   SRS308881
## 10  SRS308882
```

colData(parathyroidGenesSE)[c(21:27),]

```
## DataFrame with 7 rows and 8 columns
##            run experiment  patient treatment      time submission
study
##    <character>    <factor> <factor>  <factor> <factor>   <factor>
<factor>
## 1   SRR479072   SRX140521        4    Control      48h   SRA051611
SRP012167
## 2   SRR479073   SRX140522        4    DPN          24h   SRA051611
SRP012167
## 3   SRR479074   SRX140523        4    DPN          48h   SRA051611
SRP012167
## 4   SRR479075   SRX140523        4    DPN          48h   SRA051611
SRP012167
## 5   SRR479076   SRX140524        4    OHT          24h   SRA051611
SRP012167
## 6   SRR479077   SRX140525        4    OHT          48h   SRA051611
SRP012167
## 7   SRR479078   SRX140525        4    OHT          48h   SRA051611
SRP012167
```

```
##     sample
##    <factor>
## 1 SRS308883
## 2 SRS308884
## 3 SRS308885
## 4 SRS308885
## 5 SRS308886
## 6 SRS308887
## 7 SRS308887
```

colData(parathyroidGenesSE)[c(2,3,5,7,14,15,17,19,25,26),]

```
## DataFrame with 10 rows and 8 columns
##              run experiment  patient treatment     time submission
study
##     <character>  <factor> <factor>  <factor> <factor>    <factor>
<factor>
## 1     SRR479053  SRX140504        1   Control      48h  SRA051611
SRP012167
## 2     SRR479054  SRX140505        1       DPN      24h  SRA051611
SRP012167
## 3     SRR479056  SRX140507        1       OHT      24h  SRA051611
SRP012167
## 4     SRR479058  SRX140509        2   Control      24h  SRA051611
SRP012167
## 5     SRR479065  SRX140514        2       OHT      48h  SRA051611
SRP012167
## 6     SRR479066  SRX140515        3   Control      24h  SRA051611
SRP012167
## 7     SRR479068  SRX140517        3       DPN      24h  SRA051611
SRP012167
## 8     SRR479070  SRX140519        3       OHT      24h  SRA051611
SRP012167
## 9     SRR479076  SRX140524        4       OHT      24h  SRA051611
SRP012167
## 10    SRR479077  SRX140525        4       OHT      48h  SRA051611
SRP012167
##        sample
##       <factor>
## 1   SRS308866
## 2   SRS308867
## 3   SRS308869
## 4   SRS308871
## 5   SRS308876
## 6   SRS308877
## 7   SRS308879
## 8   SRS308881
```

```
## 9    SRS308886
## 10 SRS308887

sampletable<- parathyroidGenesSE[ ,c(2,3,5,7,14,15,17,19,25,26)]
colData(sampletable)

## DataFrame with 10 rows and 8 columns
##           run experiment  patient treatment     time submission
study
##     <character>   <factor> <factor>  <factor> <factor>   <factor>
<factor>
## 1    SRR479053  SRX140504        1   Control      48h  SRA051611
SRP012167
## 2    SRR479054  SRX140505        1   DPN          24h  SRA051611
SRP012167
## 3    SRR479056  SRX140507        1   OHT          24h  SRA051611
SRP012167
## 4    SRR479058  SRX140509        2   Control      24h  SRA051611
SRP012167
## 5    SRR479065  SRX140514        2   OHT          48h  SRA051611
SRP012167
## 6    SRR479066  SRX140515        3   Control      24h  SRA051611
SRP012167
## 7    SRR479068  SRX140517        3   DPN          24h  SRA051611
SRP012167
## 8    SRR479070  SRX140519        3   OHT          24h  SRA051611
SRP012167
## 9    SRR479076  SRX140524        4   OHT          24h  SRA051611
SRP012167
## 10   SRR479077  SRX140525        4   OHT          48h  SRA051611
SRP012167
##        sample
##      <factor>
## 1    SRS308866
## 2    SRS308867
## 3    SRS308869
## 4    SRS308871
## 5    SRS308876
## 6    SRS308877
## 7    SRS308879
## 8    SRS308881
## 9    SRS308886
## 10 SRS308887

sampletable$'Comparison'<-
as.factor(c("a","a","a","a","a","b","b","b","b","b"))
colData(sampletable)
```

```
## DataFrame with 10 rows and 9 columns
##            run experiment  patient treatment     time submission
study
##     <character>    <factor> <factor>  <factor> <factor>    <factor>
<factor>
## 1    SRR479053  SRX140504        1   Control      48h  SRA051611
SRP012167
## 2    SRR479054  SRX140505        1   DPN          24h  SRA051611
SRP012167
## 3    SRR479056  SRX140507        1   OHT          24h  SRA051611
SRP012167
## 4    SRR479058  SRX140509        2   Control      24h  SRA051611
SRP012167
## 5    SRR479065  SRX140514        2   OHT          48h  SRA051611
SRP012167
## 6    SRR479066  SRX140515        3   Control      24h  SRA051611
SRP012167
## 7    SRR479068  SRX140517        3   DPN          24h  SRA051611
SRP012167
## 8    SRR479070  SRX140519        3   OHT          24h  SRA051611
SRP012167
## 9    SRR479076  SRX140524        4   OHT          24h  SRA051611
SRP012167
## 10   SRR479077  SRX140525        4   OHT          48h  SRA051611
SRP012167
##         sample Comparison
##      <factor>   <factor>
## 1  SRS308866          a
## 2  SRS308867          a
## 3  SRS308869          a
## 4  SRS308871          a
## 5  SRS308876          a
## 6  SRS308877          b
## 7  SRS308879          b
## 8  SRS308881          b
## 9  SRS308886          b
## 10 SRS308887          b
```

```r
library("DESeq2")
se<- sampletable
dds<- DESeqDataSet(se, design = ~ treatment + Comparison)
dds
```

```
## class: DESeqDataSet
## dim: 63193 10
## metadata(2): MIAME version
## assays(1): counts
```

```
## rownames(63193): ENSG00000000003 ENSG00000000005 ... LRG_98 LRG_99
## rowData names(0):
## colnames: NULL
## colData names(9): run experiment ... sample Comparison

keep <- rowSums(counts(dds)) > 1
dds <- dds[keep,]
nrow(dds)

## [1] 27077

rld<-rlog(dds,blind=F)
rld

## class: DESeqTransform
## dim: 27077 10
## metadata(2): MIAME version
## assays(1): ''
## rownames(27077): ENSG00000000003 ENSG00000000005 ...
ENSG00000271707
##     ENSG00000271711
## rowData names(7): baseMean baseVar ... dispFit rlogIntercept
## colnames(10): 1 2 ... 9 10
## colData names(10): run experiment ... Comparison sizeFactor

library("gplots")

##
## Attaching package: 'gplots'

## The following object is masked from 'package:IRanges':
##
##     space

## The following object is masked from 'package:S4Vectors':
##
##     space

## The following object is masked from 'package:stats':
##
##     lowess

topVarGenes <- head(order(-rowVars(assay(rld))),100)
topVarGenes

##   [1]  1754 15037  2960 15449  4629 13495  2727  7271  2295 12034
730  6214
##  [13] 10950 14824  2151  8218 10460  2742  5883 15035  8836  3816
2532  9380
##  [25] 15315  5504  2490  4444  9132 20937 10249  1938  8331  7125
```

```
9799  5810
##  [37] 15047 12185 14263  8530  9167  3205 15778 11971  6257 10263
2888  6309
##  [49]  1258  3961  4557  2117  6288 10917 11229  5662    41  2718
15397  6356
##  [61]  3348 20635  2687  9586 14992 22041   623 14731  1473  6529
3792 14139
##  [73] 14314 15413   393 12552  3414  5710    35  2415  6264 10719
5899 16582
##  [85]  5854 13745 13844 11584  3237  3861  9382 11623  8057  2026
1677  8696
##  [97]  8084   388 21144  8326
```

```
heatmap.2(x = assay(rld)[topVarGenes,])
```
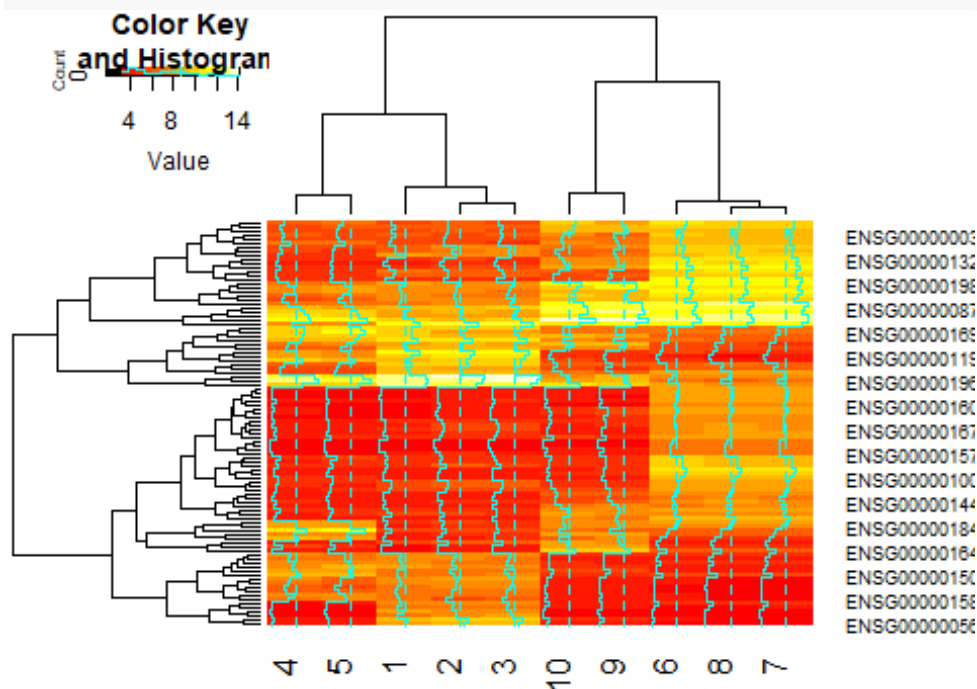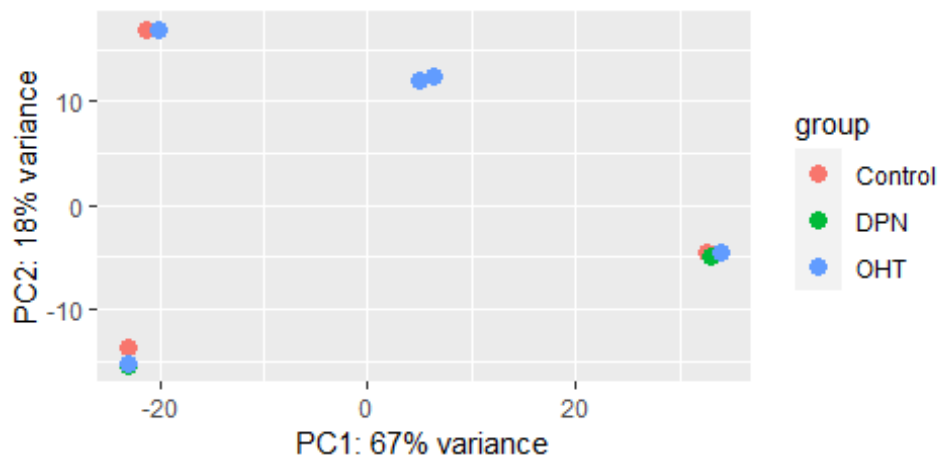


```
library("ggfortify")
```

```
## Loading required package: ggplot2
```

```
plotPCA(rld, intgroup = c("treatment"))
```

```
dds <- DESeq(dds)

## estimating size factors

## estimating dispersions

## gene-wise dispersion estimates

## mean-dispersion relationship

## final dispersion estimates

## fitting model and testing

resultsNames(dds)

## [1] "Intercept"                  "treatment_DPN_vs_Control"
## [3] "treatment_OHT_vs_Control"   "Comparison_b_vs_a"

res <- results(dds, contrast = c("Comparison", "a", "b"))
res1<- subset(res, log2FoldChange > 2 & padj < 0.05)
res2 <- head(res1, 10)
summary(res2)

##
## out of 10 with nonzero total read count
## adjusted p-value < 0.1
## LFC > 0 (up)        : 10, 100%
```

```
## LFC < 0 (down)       : 0, 0%
## outliers [1]         : 0, 0%
## low counts [2]       : 0, 0%
## (mean count < 2)
## [1] see 'cooksCutoff' argument of ?results
## [2] see 'independentFiltering' argument of ?results

library("EnhancedVolcano")

## Loading required package: ggrepel

require("ggrepel")
EnhancedVolcano(res2,
                lab = rownames(res2),
                x = 'log2FoldChange',
                y = 'pvalue',)

## Warning: Removed 1 rows containing missing values (`geom_vline()`).
```
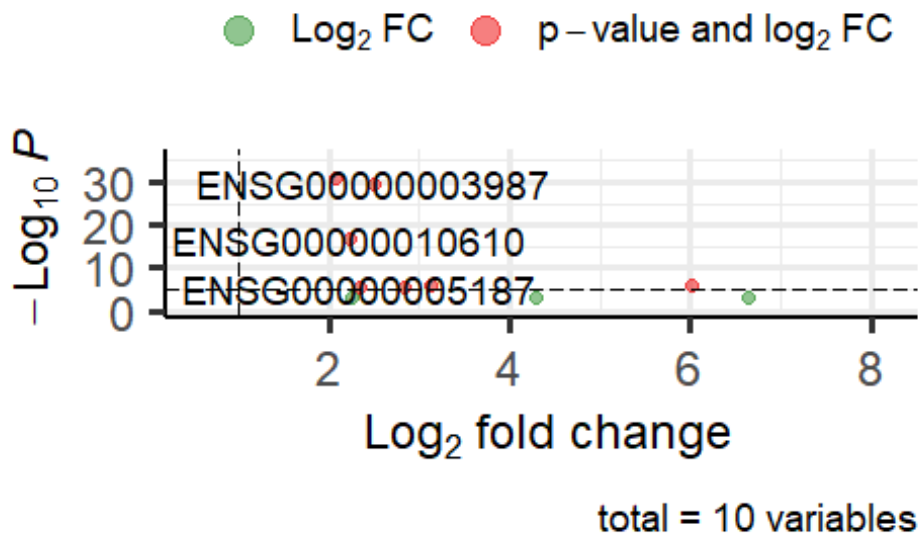


```
mydata <- read.csv("C:\\Users\\Divya\\Desktop\\enrichment.csv")
mydata

##    Enrichment.FDR nGenes Pathway.Genes Fold.Enrichment
## 1      0.03186634      2          1527        7.464309
## 2      0.03010828      1            73       78.068493
## 3      0.03010828      1            28      203.535714
```

```
## 4      0.03010828      1       97       58.752577
## 5      0.03010828      1       78       73.064103
## 6      0.03010828      1       99       57.565657
## 7      0.03010828      1       92       61.945652
## 8      0.03010828      1      108       52.768519
## 9      0.03010828      1      103       55.330097
## 10     0.03010828      1       89       64.033708
## 11     0.03010828      1       38      149.973684
## 12     0.03186634      1      149       38.248322
## 13     0.03186634      1      137       41.598540
## 14     0.04095076      1      222       25.671171
##                                                          Pathway
## 1                                              Metabolic pathways
## 2                                  Inositol phosphate metabolism
## 3                                            Butanoate metabolism
## 4                        Phosphatidylinositol signaling system
## 5                              Antigen processing and presentation
## 6                                    Hematopoietic cell lineage
## 7                              Th1 and Th2 cell differentiation
## 8                                      Th17 cell differentiation
## 9                              T cell receptor signaling pathway
## 10 PD-L1 expression and PD-1 checkpoint pathway in cancer
## 11                                   Primary immunodeficiency
## 12                                     Cell adhesion molecules
## 13                                          Yersinia infection
## 14                     Human T-cell leukemia virus 1 infection
##                                                              URL
Genes
## 1  http://www.genome.jp/kegg-bin/show_pathway?hsa01100    MTMR7
ACSM3
## 2  http://www.genome.jp/kegg-bin/show_pathway?hsa00562
MTMR7
## 3  http://www.genome.jp/kegg-bin/show_pathway?hsa00650
ACSM3
## 4  http://www.genome.jp/kegg-bin/show_pathway?hsa04070
MTMR7
## 5  http://www.genome.jp/kegg-bin/show_pathway?hsa04612
CD4
## 6  http://www.genome.jp/kegg-bin/show_pathway?hsa04640
CD4
## 7  http://www.genome.jp/kegg-bin/show_pathway?hsa04658
CD4
## 8  http://www.genome.jp/kegg-bin/show_pathway?hsa04659
CD4
## 9  http://www.genome.jp/kegg-bin/show_pathway?hsa04660
CD4
```

```
## 10 http://www.genome.jp/kegg-bin/show_pathway?hsa05235
CD4
## 11 http://www.genome.jp/kegg-bin/show_pathway?hsa05340
CD4
## 12 http://www.genome.jp/kegg-bin/show_pathway?hsa04514
CD4
## 13 http://www.genome.jp/kegg-bin/show_pathway?hsa05135
CD4
## 14 http://www.genome.jp/kegg-bin/show_pathway?hsa05166
CD4
```

7. If there are differentially expressed genes and enriched pathways, explain how the original treatment and factor conditions of the samples you compared could have effected this or if there aren't, justify why there are no differentially expressed genes?

Instead of differentially expressed genes and enriched pathways, we have compared with our sample column comparison to treatment which must have varied the results. The results would have been different if we used original treatment and factor conditions.

2. Difficulty Adjustment

• How long did this assignment take you to complete?

It took us 48 hours to finish this assignment.

• If the assignment took you longer than the 10 hours, which parts were overly difficult?

We had difficulty in installing the packages and also learning R was challenging as we had to start from scratch.