# Methodology

Step 1 : Call Library (edgeR)

*‘library(edgeR)’*

Step 2 : Load in Dataset into count\_data variable and split Dataset into sample set by dividing # of samples in half unless otherwise specified.

*‘count\_data <- read.table("DATASETNAME.tsv", header=TRUE, row.names=1)*

*groups <- factor(rep(1:2, each=NUMBEROFCOLS/2))’*

Step 3 : Create DGEList Object (kind of like a specific data array for this library)

*‘y <- DGEList(counts=count\_data, group=groups)’*

Step 4 : ***Optional*** – Filter out low count genes to minimise skew

*‘keep <- filterByExpr(y)*

*y <- y[keep, ]’*

Step 5 : Normalise Values in dataset

‘y <- calcNormFactors(y)’

Step 6 : Initialise Matrix for Values (2 group comparisons or more)

*‘design <- model.matrix(~groups)’*

Step 7 : EdgeR Dispersion test

*‘y <- estimateDisp(y, design)’*

Step 8 : Extract Differential Expression data

*‘et <- exactTest(y)’*

Step 9: Summarise with DecideTestsDGE

*‘decided <- decideTestsDGE(et, adjust.method="BH", p.value=0.05)*

*summary(decided)’*

Step 10 : Print output of top genes

*‘topTags(et)’*

Step 11 : Load in Metadata and join to unfiltered TopTags Variable

*‘meta\_data <- read.table("FILENAME\_meta.tsv", header=TRUE, row.names=1)*

*edgeR\_results <- topTags(et, n=Inf)*

*annotated\_results <- merge(edgeR\_results, meta\_data, by="row.names", all.x=TRUE)*

*rownames(annotated\_results) <- annotated\_results$Row.names*

*annotated\_results$Row.names <- NULL*

Step 12 : Comparison using LogFC (can change this)

*# Up Regulated :*

*‘detected\_up\_edgeR <- rownames(annotated\_results[annotated\_results$logFC > 0 & annotated\_results$FDR < 0.05,])*

*meta\_up <- rownames(annotated\_results[annotated\_results$upregulation == 1,])*

*common\_up <- intersect(detected\_up\_edgeR, meta\_up)’*

# *Down Regulated :*

*detected\_down\_edgeR <- rownames(annotated\_results[annotated\_results$logFC < 0 & annotated\_results$FDR < 0.05,])*

*meta\_down <- rownames(annotated\_results[annotated\_results$downregulation == 1,])*

*common\_down <- intersect(detected\_down\_edgeR, meta\_down)*

Step 13 : Summarise outliers

*outliers\_up <- setdiff(detected\_up\_edgeR, meta\_up)*

*write.csv(outliers\_up, "FILENAME\_outliers\_upregulated.csv", row.names = FALSE)*

*outliers\_down <- setdiff(detected\_down\_edgeR, meta\_down)*

*write.csv(outliers\_down, "FILENAME\_outliers\_downregulated.csv", row.names = FALSE)*

Step 14 : Accuracy and Precision Matrix

*true\_positives <- length(common\_up) + length(common\_down)*

*false\_positives <- length(setdiff(detected\_up\_edgeR, meta\_up)) + length(setdiff(detected\_down\_edgeR, meta\_down))*

*true\_negatives <- nrow(meta\_data) - (length(meta\_up) + length(meta\_down)) - false\_positives*

*false\_negatives <- length(setdiff(meta\_up, detected\_up\_edgeR)) + length(setdiff(meta\_down, detected\_down\_edgeR))*

*accuracy <- (true\_positives + true\_negatives) / nrow(meta\_data)*

*precision <- true\_positives / (true\_positives + false\_positives)*

*recall <- true\_positives / (true\_positives + false\_negatives)*

f1\_score <- 2 \* ((precision \* recall) / (precision + recall))

Step 15 : Venn Diagram Creation

*Library(VennDiagram)*

*# upregulated Genes*

*venn.diagram(*

*x = list(edgeR = detected\_up\_edgeR, meta = meta\_up),*

*category.names = c("edgeR detected up", "Metadata up"),*

*output = TRUE,*

*filename = "FILENAME\_venn\_upregulated.png",*

*output.type = "png",*

*imagetype = "png",*

*resolution = 300*

*)*

*# Venn diagram for downregulated genes*

*venn.diagram(*

*x = list(edgeR = detected\_down\_edgeR, meta = meta\_down),*

*category.names = c("edgeR detected down", "Metadata down"),*

*output = TRUE,*

*filename = "FILENAME\_venn\_downregulated.png",*

*output.type = "png",*

*imagetype = "png",*

*resolution = 300*

*)*

*Link to Code used :*

**

**

Accuracy : True Positives + True Negatives / Total genes

Precision : True Positives / True positives + False Positives

Recall : True Positives / True Positives + False Negatives

F1 Score : Mean of precision and Recall

# Dataset 3\_500\_500

Sample Size : 6 samples across 9987 entries

Total Dif Expression : 240

Total Upregulated : 123

Total Downregulated : 118

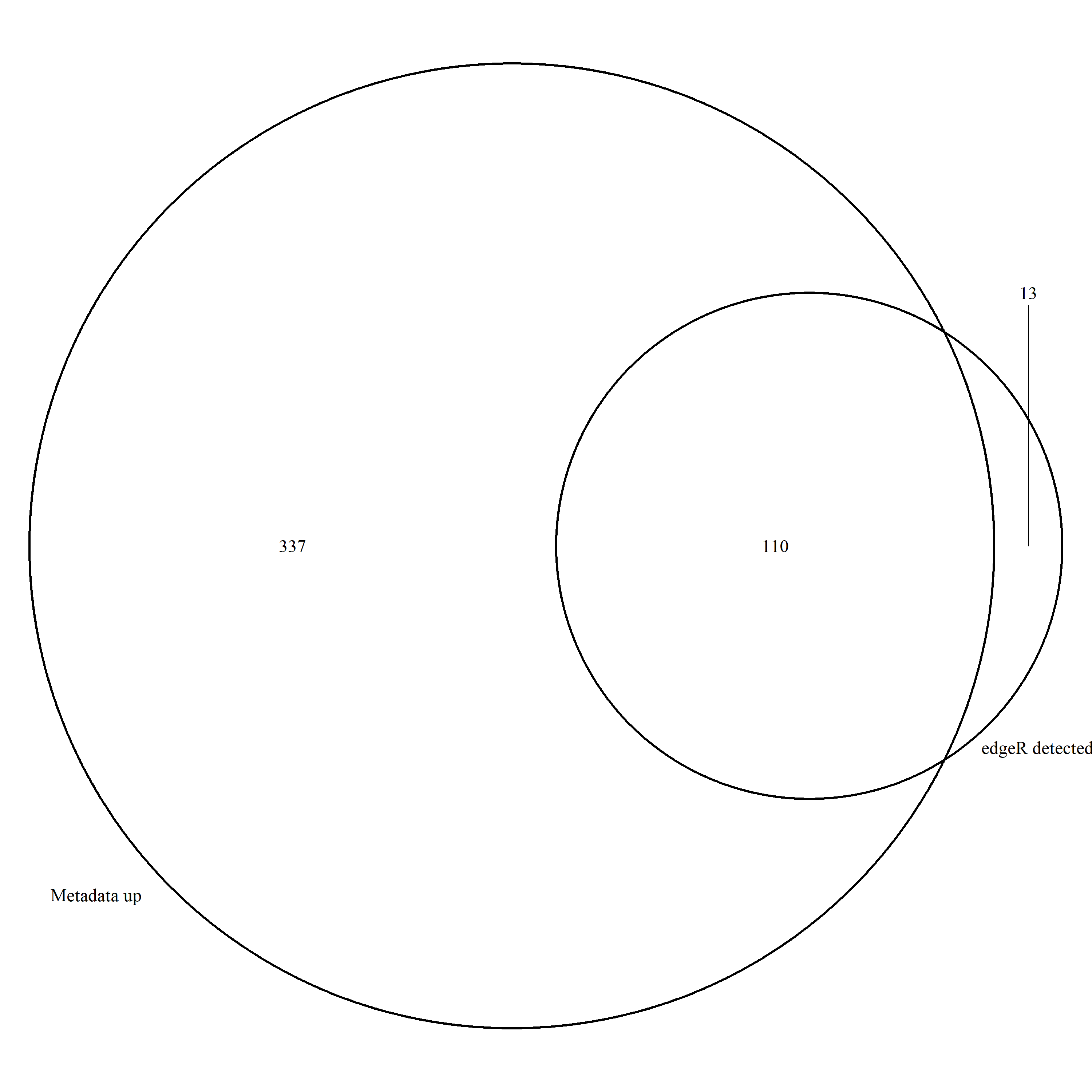
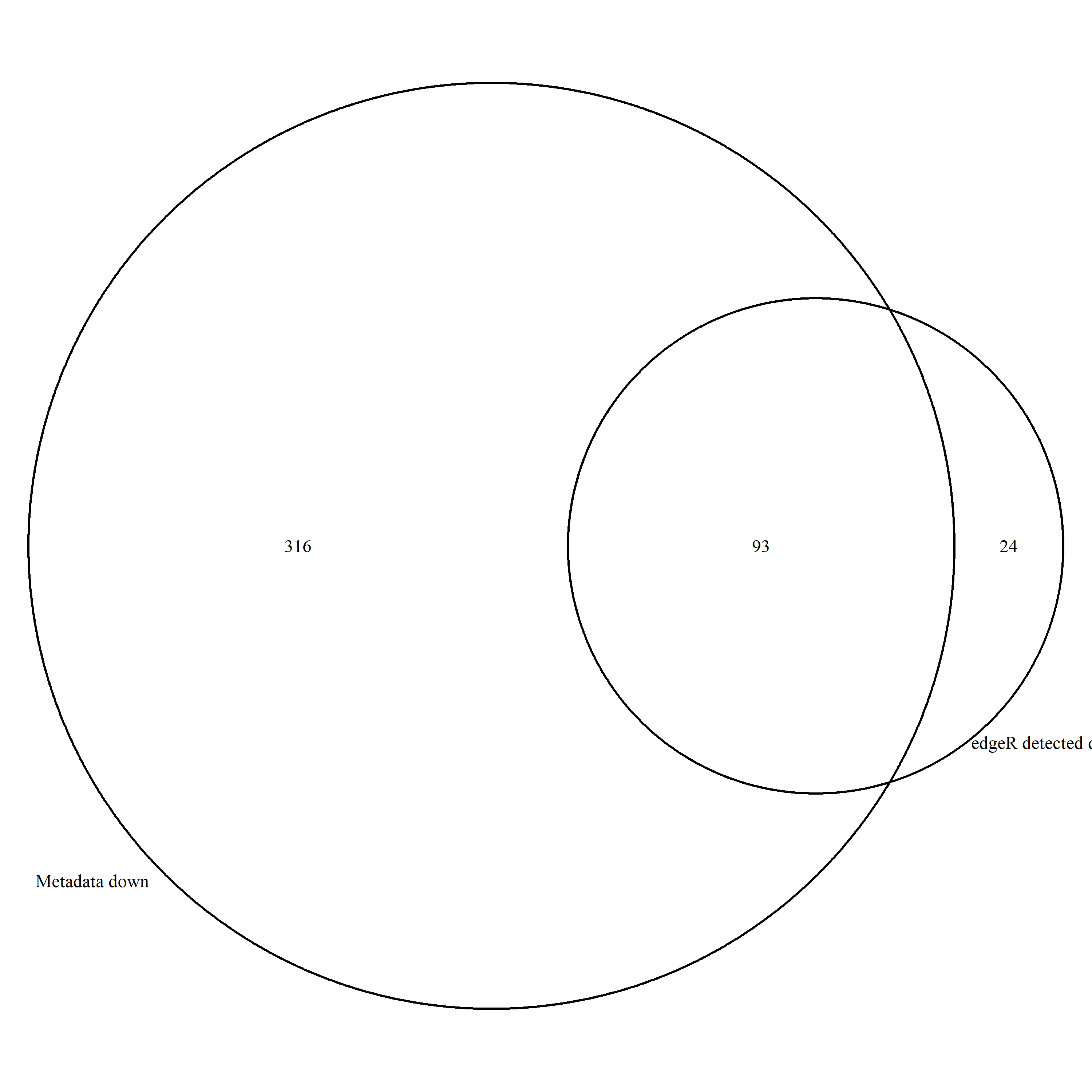


Figure Down Regulated

Figure Up regulated

**Accuracy and Precision Evaluation**

Accuracy : .93

Precision : .84

Recall : .23

F1 Score : 0.37

**Evaluation Metrics**

False Discovery Rate : Threshold is 0.05

P-Vaule Mean :

MetaData

Edge R Upregulated

EdgeR Downregulated

**Consistency**

Genes that are differentially expressed across down and up regulated methods

**Outliers:**

UpReg g1395 g1770 g3775 g3797 g4167 g5711 g5840 g6570 g6696 g7568 g7590 g8686 g9085

DownReg g1322 g3341 g3731 g3741 g3809 g3867 g4255 g4354 g4631 g4658 g6133 g6211 g6709 g6801 g6813 g7139 g7450 g7489 g7944 g7957 g8481 g8534 g9032 g9078

# Dataset 6\_750\_250

Sample Size : 12 samples across 9997 entries

Total Dif Expression : 468

Total Upregulated : 260

Total Downregulated : 208

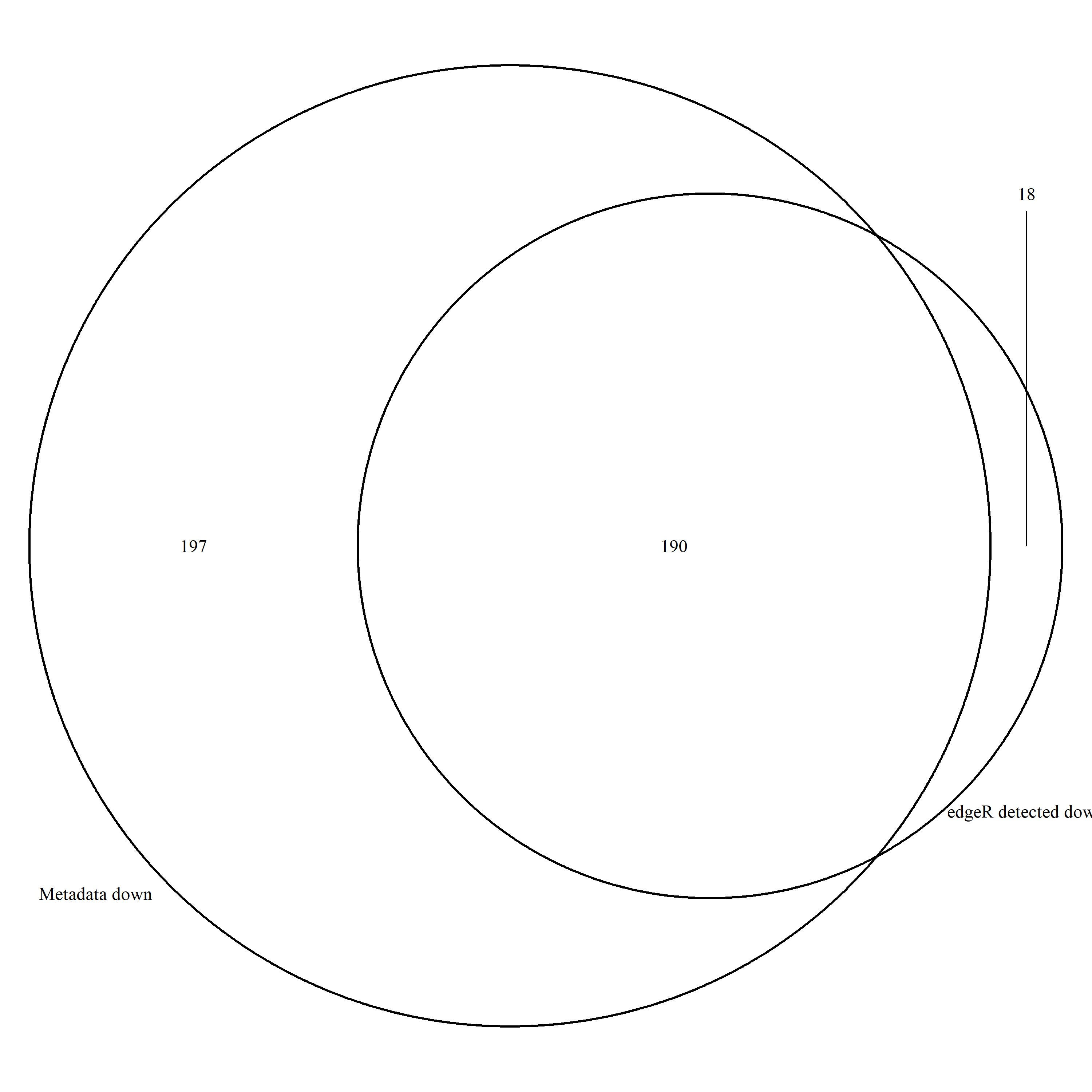
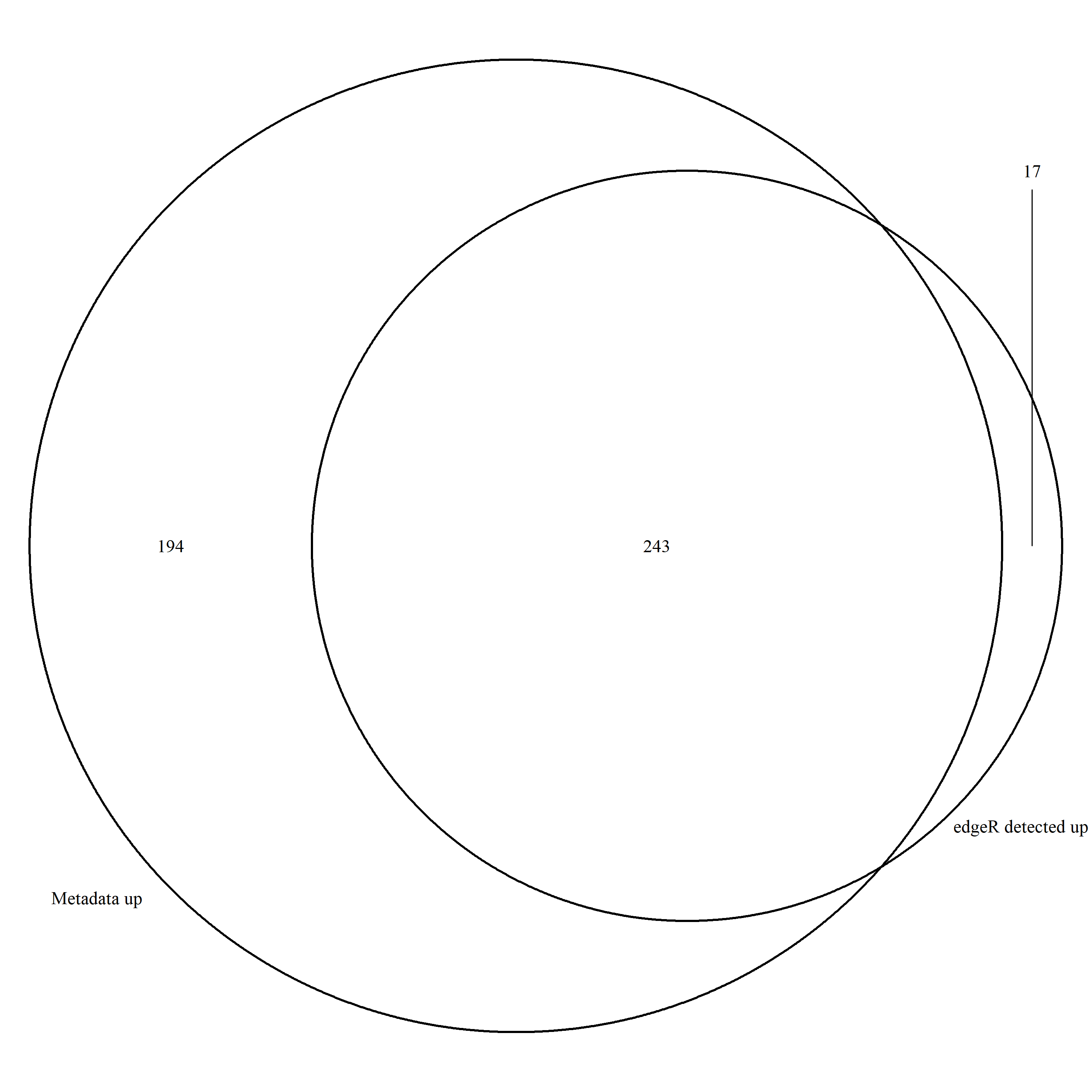


Figure - Upregulated

Figure - Downregulated

**Accuracy and Precision Evaluation**

Accuracy : 0.957

Precision : 0.9252

Recall : 0.5254

F1 Score : 0.6702

**Evaluation Metrics**

False Discovery Rate : Threshold is 0.05

P-Vaule Mean :

MetaData

Edge R Upregulated

EdgeR Downregulated

**Consistency**

Genes that are differentially expressed across down and up regulated methods

**Outliers:**

UpReg g2116 g2219 g2325 g3016 g3555 g3660 g4065 g4436 g4730 g6400 g6461 g7579 g7595 g7700 g7958 g8933

DownReg g1309 g2007 g3255 g3537 g4402 g4440 g4473 g4545 g4738 g4833 g6551 g7393 g7457 g7462 g8174 g8818 g9224 g9546 g9773 g9898

# DataSet 9\_750\_250

Sample Size : 18 samples across 9999 entries

Total Dif Expression : 611

Total Upregulated : 442

Total Downregulated : 169

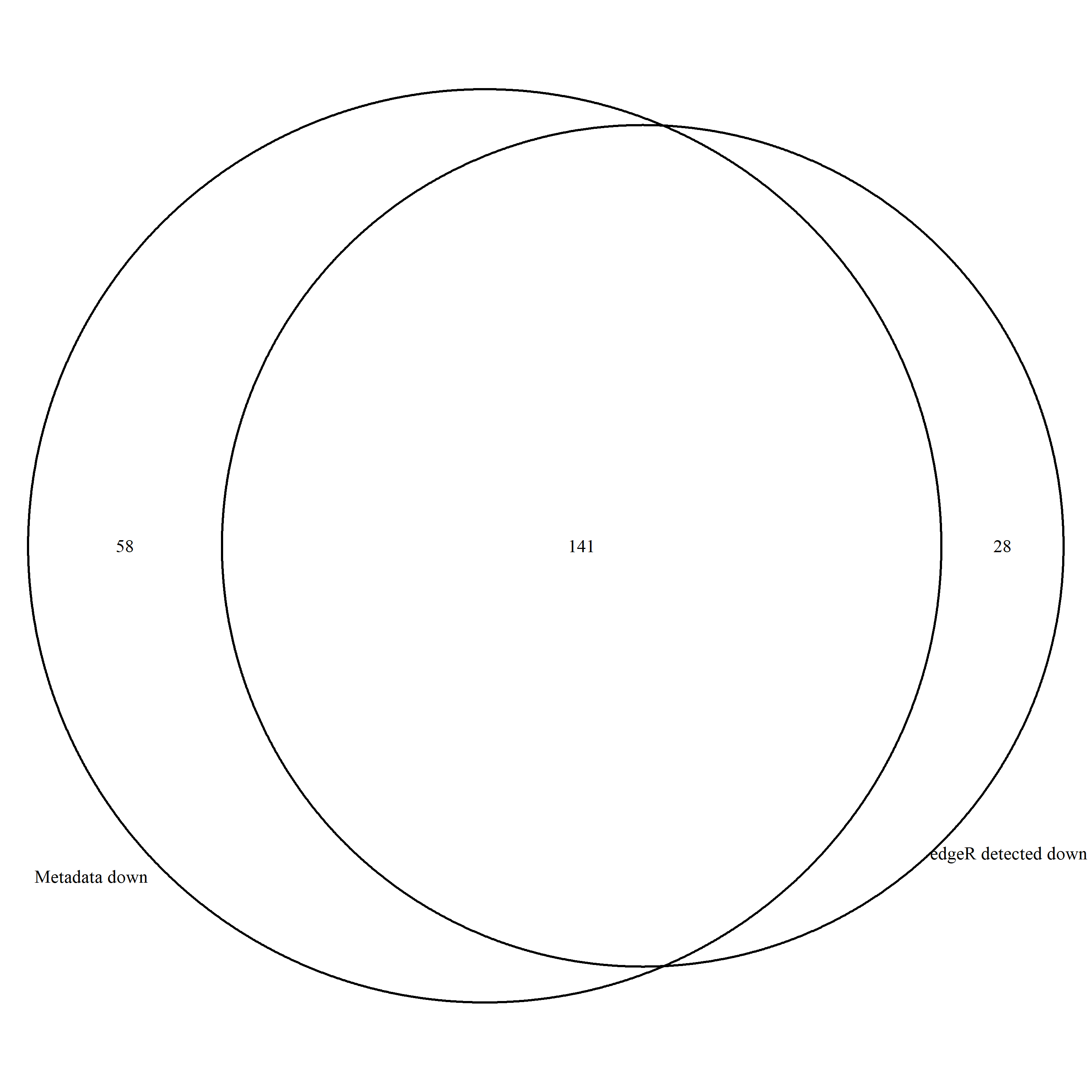
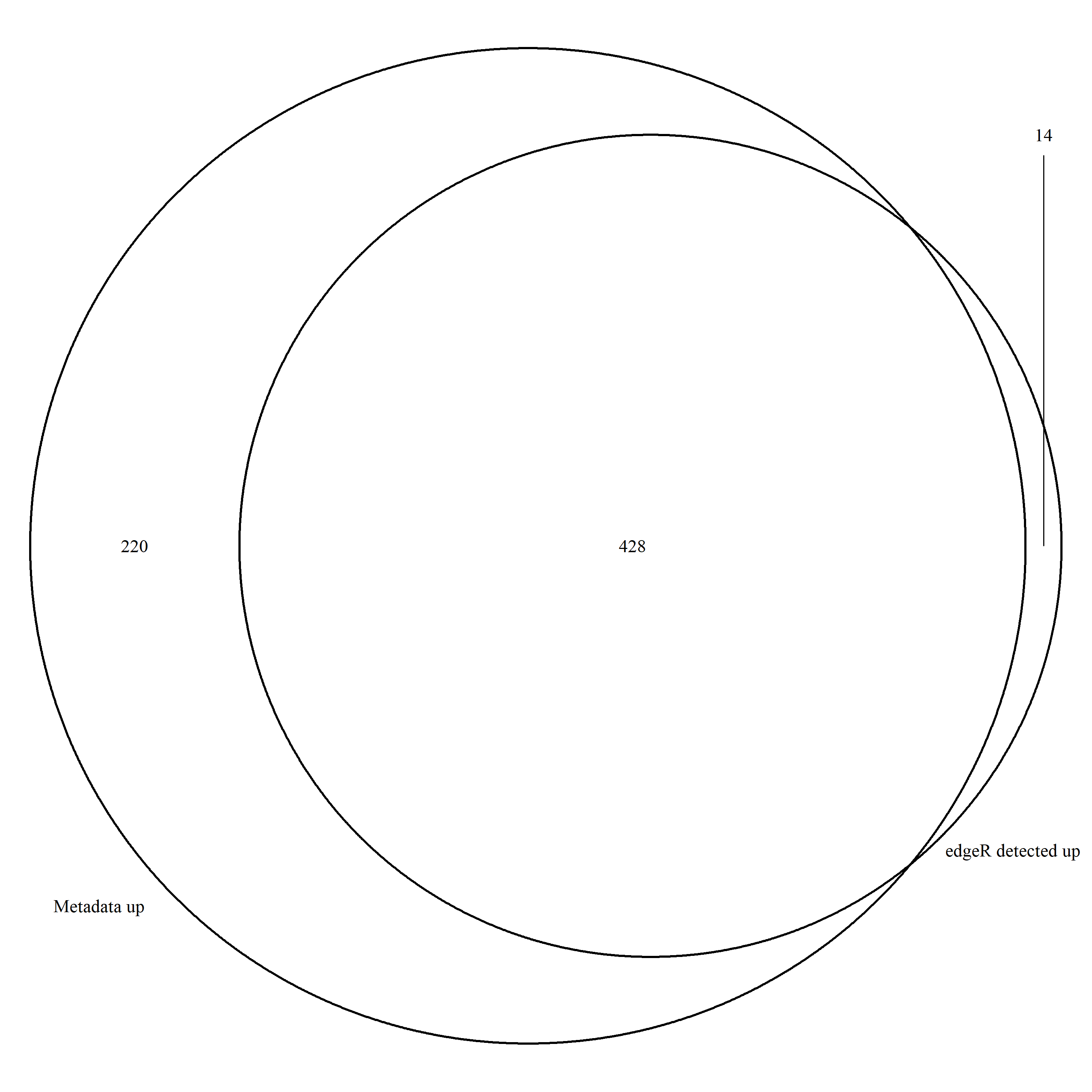


Figure - Upregulated

Figure - Downregulated

**Accuracy and Precision Evaluation**

Accuracy : 0.967

Precision : 0.931

Recall : 0.6717

F1 Score : 0.78

**Evaluation Metrics**

False Discovery Rate : Threshold is 0.05

P-Vaule Mean :

MetaData

Edge R Upregulated

EdgeR Downregulated

**Consistency**

Genes that are differentially expressed across down and up regulated methods

**Outliers:**

UpRegulated : g1162 g1286 g1608 g2795 g4236 g4481 g4965 g5127 g5335 g6038 g6381 g7952 g8694 g9803

DownRegulated : g1094 g1932 g2015 g2088 g2595 g2880 g3019 g3771 g3818 g4095 g4412 g4973 g5243 g5486 g5588 g5844 g5950 g6162 g6382 g6676 g7130 g7400 g7914 g8353 g8369 g8794 g9186 g9612

# Dataset 9\_1000\_0

Sample Size : 18 samples across 9999 entries

Total Dif Expression :

Total Upregulated :

Total Downregulated :

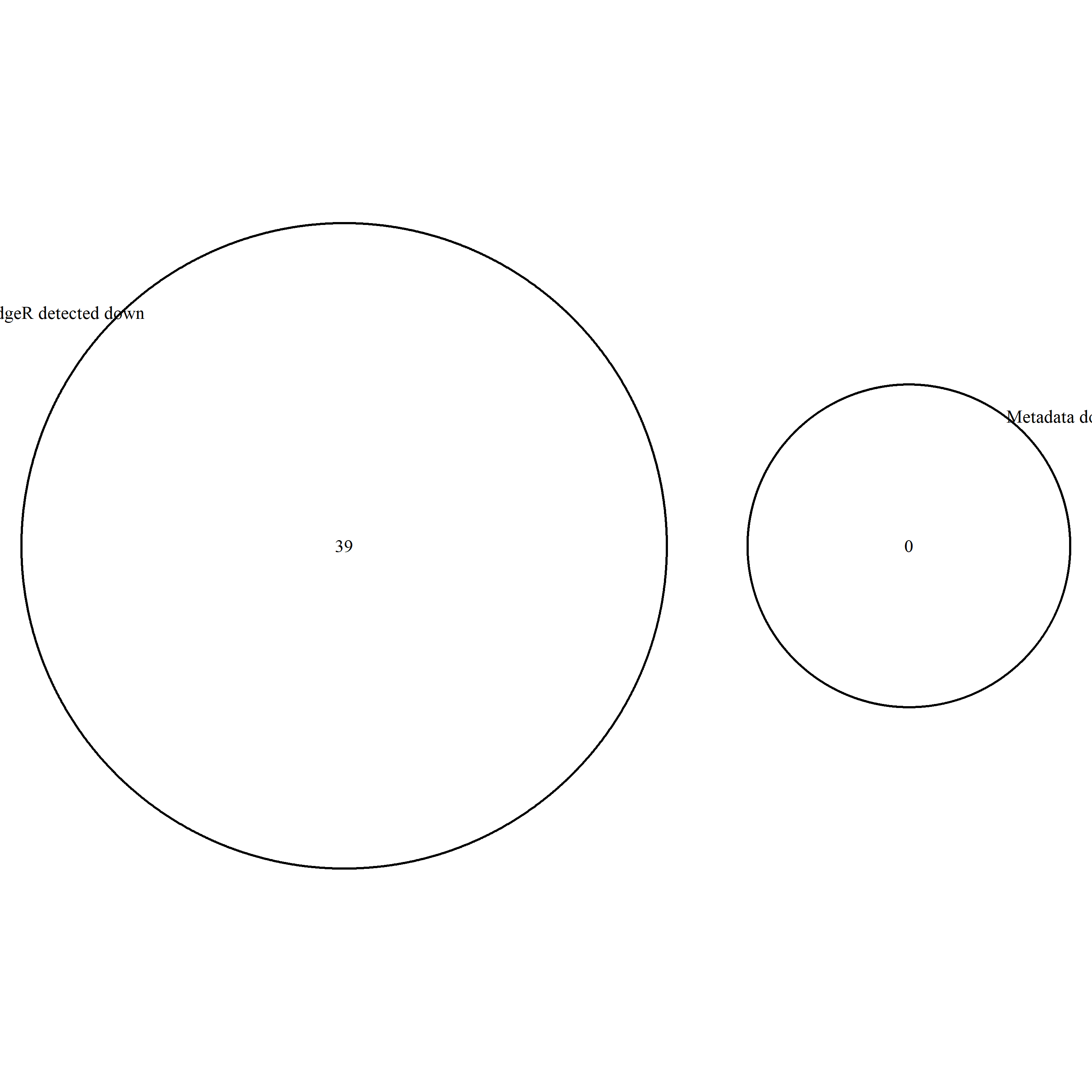
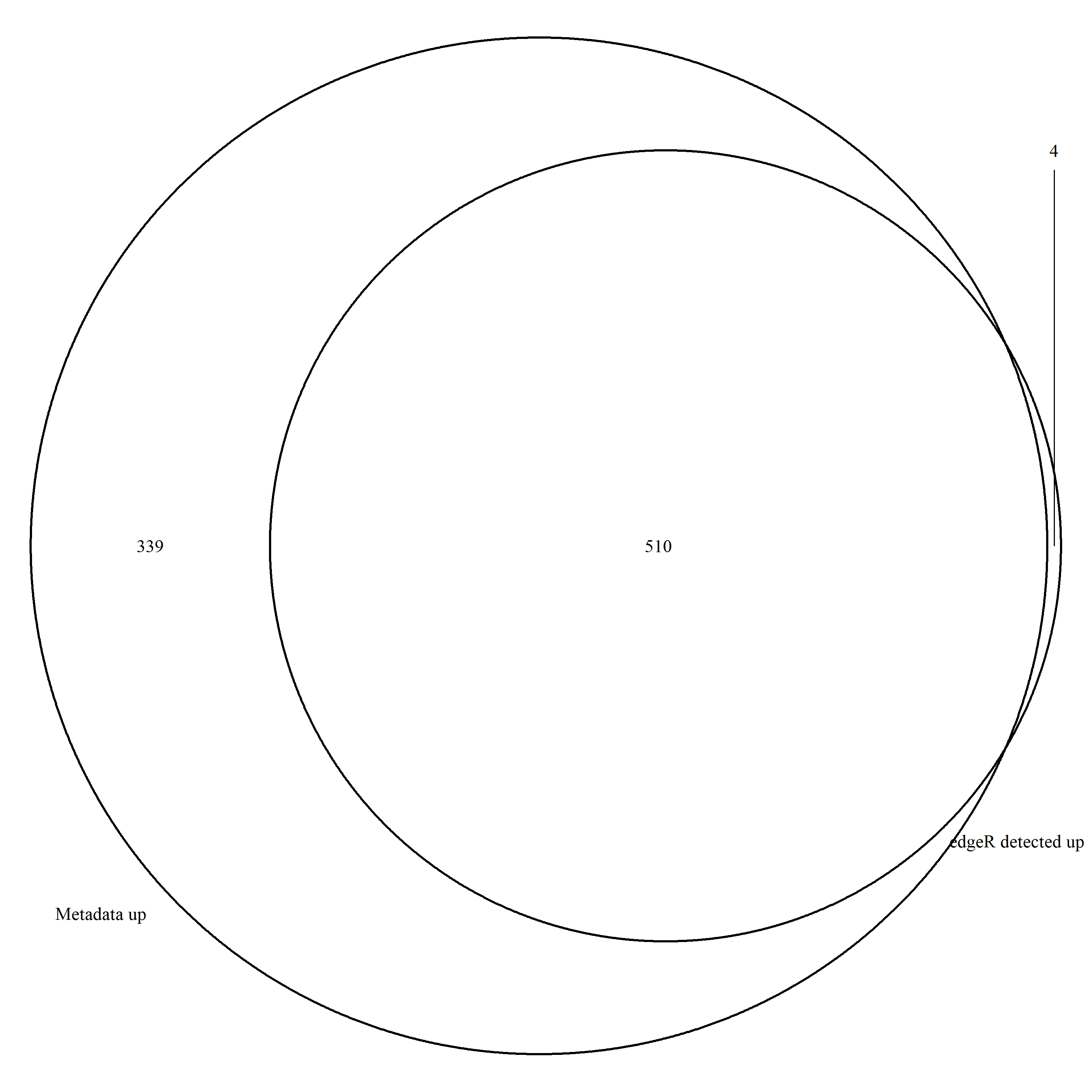


Figure - Upregulated

Figure - Downregulated

**Accuracy and Precision Evaluation**

Accuracy : 0.961

Precision : 0.922

Recall : 0.600

F1 Score : 0.72

**Evaluation Metrics**

False Discovery Rate : Threshold is 0.05

**Consistency**

**Outliers:**

UpRegulated : g1746 g3067 g4735 g7580

DownRegulated : g1045 g1289 g1370 g1744 g1900 g2156 g2190 g2524 g2612 g2695 g2964 g3456 g3888 g4334 g4399 g4522 g4847 g4865 g5241 g5319 g5413 g5440 g5700 g5901 g6841 g6977 g7176 g7504 g7573 g7601 g7702 g8409 g8491 g8704 g8982 g9156 g9256 g9283 g9975

# Next Steps :

-We run similar code against the same 4 datasets for DESeq2 and NOISeq

- Identify similarly expressed genes for each test (what’s consistently up regulated for all methods vs the metadata baseline)

- For each tool how impacted are they for parameters used? For instance, edgeR is impacted greatly by the False Discovery Threshold, what parameters are similarly impactful for DESeq2 and NOISeq?

- What does this tell us ?

For EdgeR, because its an empirical Bayes estimation its pretty good at not being skewed by outliers but is weak against datasets that are differentially expressed .

For NOISeq it’s a bit slower because of the computational method employed but is more resilient to noisy data when outputting its results

# Report Format :

## Introduction

* What is differential expression analysis?
* Why is it important in specific reference to RNA sequencing and analysis?

## Literature Review :

### DESeq2

Zhang, Z. H., Jhaveri, D. J., Marshall, V. M., Bauer, D. C., Edson, J., Narayanan, R. K., ... & Zhao, Q. Y. (2014). A comparative study of techniques for differential expression analysis on RNA-Seq data. *PloS one*, *9*(8), e103207.

### NOISeq :

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4666377/>

<https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-14-S13-S7>

### EdgeR

<https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0157022>

<https://pubmed.ncbi.nlm.nih.gov/19910308/>

<https://pubmed.ncbi.nlm.nih.gov/34605806/>

<https://pubmed.ncbi.nlm.nih.gov/27280887/>

## Project Definition / Research Question :

* What is the impact of different computational packages on predicting or classifying differentially expressed genes based on a known dataset?
* What are the strengths and weaknesses of each package / tool on classification?

## Methods :

* Consistent methodology
  + Eg – Load Data
  + Normalise, extract data etc

## Results

* Consistent tables for differentially expressed genes.
* Accuracy and Precision Evaluation for each tool
* Consistent Visualisations (Venn Diagrams, Volcano Plots etc)

Eg:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Dataset 9\_1000\_0** | | | | | |
| Up Regulated | | | Down Regulated | | |
| **DESEQ2** | **EdgeR** | **NOISEQ** | **DESEQ2** | **EdgeR** | **NOISEQ** |
| g1253 | g1746 | g314 | g1045 | g1045 |  |
| g1746 | g3067 | g928 | g1230 | g1289 |  |
| g3067 | g4735 |  | g1289 | g1370 |  |
| g4735 | g7580 |  | g1292 | g1744 |  |
| g4762 |  |  | g1340 | g1900 |  |
| g5194 |  |  | g1370 | g2156 |  |
| g5983 |  |  | g1461 | g2190 |  |
| g6077 |  |  | g1727 | g2524 |  |
| g7021 |  |  | g1744 | g2612 |  |
| g7580 |  |  | g1800 | g2695 |  |
| g8114 |  |  | g1900 | g2964 |  |
| g8675 |  |  | g2190 | g3456 |  |
|  |  |  | g2287 | g3888 |  |
|  |  |  | g2340 | g4334 |  |
|  |  |  | g2524 | g4399 |  |
|  |  |  | g2612 | g4522 |  |
|  |  |  | g2695 | g4847 |  |
|  |  |  | g2821 | g4865 |  |
|  |  |  | g2825 | g5241 |  |
|  |  |  | g2906 | g5319 |  |
|  |  |  | g2964 | g5413 |  |
|  |  |  | g3456 | g5440 |  |
|  |  |  | g3493 | g5700 |  |
|  |  |  | g3571 | g5901 |  |
|  |  |  | g3591 | g6841 |  |
|  |  |  | g3888 | g6977 |  |
|  |  |  | g4251 | g7176 |  |
|  |  |  | g4334 | g7504 |  |
|  |  |  | g4399 | g7573 |  |
|  |  |  | g4522 | g7601 |  |
|  |  |  | g4733 | g7702 |  |
|  |  |  | g4847 | g8409 |  |
|  |  |  | g4865 | g8491 |  |
|  |  |  | g5060 | g8704 |  |
|  |  |  | g5241 | g8982 |  |
|  |  |  | g5319 | g9156 |  |
|  |  |  | g5413 | g9256 |  |
|  |  |  | g5440 | g9283 |  |
|  |  |  | g5480 | g9975 |  |
|  |  |  | g5632 |  |  |
|  |  |  | g5698 |  |  |
|  |  |  | g5700 |  |  |
|  |  |  | g5901 |  |  |
|  |  |  | g5905 |  |  |
|  |  |  | g6365 |  |  |
|  |  |  | g6481 |  |  |
|  |  |  | g6577 |  |  |
|  |  |  | g6841 |  |  |
|  |  |  | g6977 |  |  |
|  |  |  | g7176 |  |  |
|  |  |  | g7270 |  |  |
|  |  |  | g7341 |  |  |
|  |  |  | g7504 |  |  |
|  |  |  | g7573 |  |  |
|  |  |  | g7601 |  |  |
|  |  |  | g7702 |  |  |
|  |  |  | g7958 |  |  |
|  |  |  | g8409 |  |  |
|  |  |  | g8491 |  |  |
|  |  |  | g8676 |  |  |
|  |  |  | g8704 |  |  |
|  |  |  | g8817 |  |  |
|  |  |  | g8910 |  |  |
|  |  |  | g8982 |  |  |
|  |  |  | g9156 |  |  |
|  |  |  | g9221 |  |  |
|  |  |  | g9256 |  |  |
|  |  |  | g9261 |  |  |
|  |  |  | g9283 |  |  |
|  |  |  | g9384 |  |  |
|  |  |  | g9802 |  |  |
|  |  |  | g9975 |  |  |

## Discussion: