SVA Recount

Christopher Lo

9/19/2022

Introduction

What happens to the number of Surrogate Variables as we increase the number of studies in recount3?

To illustrate this, we examined 4 studies that involved case/control prostate tumor and normal samples, with a minimal of 25 samples per study. For each study, we examined the number of SVs (using method = "be"), and we also aggregated them one by one and examined the number of SVs.

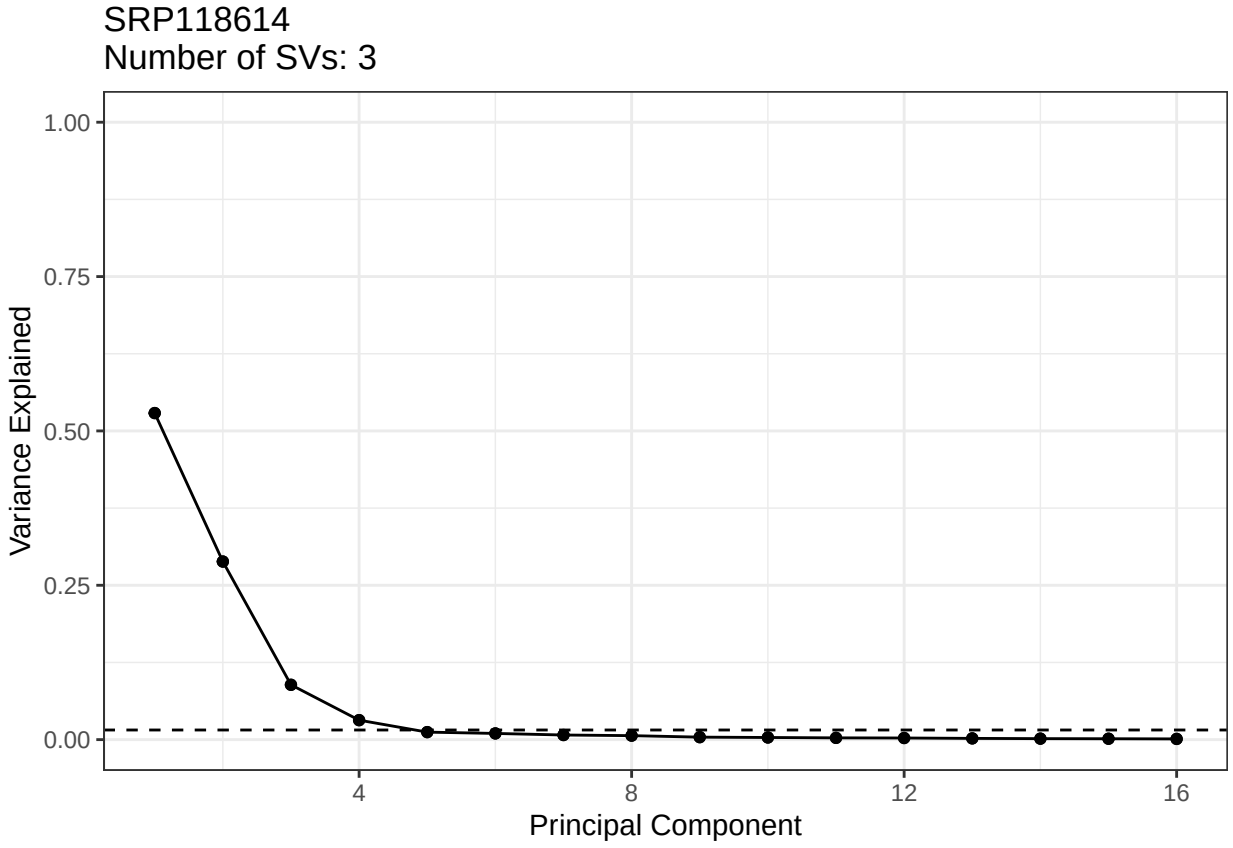
1. SRP118614: "Overall design: Matched high-grade (GS=7(4+3)) prostate tumor and adjacent normal specimens from 16 patients (8 AAM and 8 EAM) were subjected to two replicate runs of RNA-sequencing."
2. SRP002628: "Overall design: We sequenced the transcriptome (polyA+) of 20 prostate cancer tumors and 10 matched normal tissues using Illumina GAII platform. Then we used bioinformatic approaches to identify prostate cancer specific aberrations which include gene fusion, alternative splicing, somatic mutation."
3. SRP212704: "Overall design: Strand specific total RNA seq was performed using frozen patient matched prostate cancer tissue in biological duplicates. Purpose: The goal of present study is to compare transcript level changes between normal and tumor of same individuals"
4. SRP027258: "We utilized RNA sequencing to test the hypothesis that SFN modifies the expression of genes that are critical in prostate cancer progression. Normal prostate epithelial cells, and androgen-dependent and androgen-independent prostate cancer cells were treated with 15 µM SFN and the transcriptome was determined at 6 and 24 hour time points."

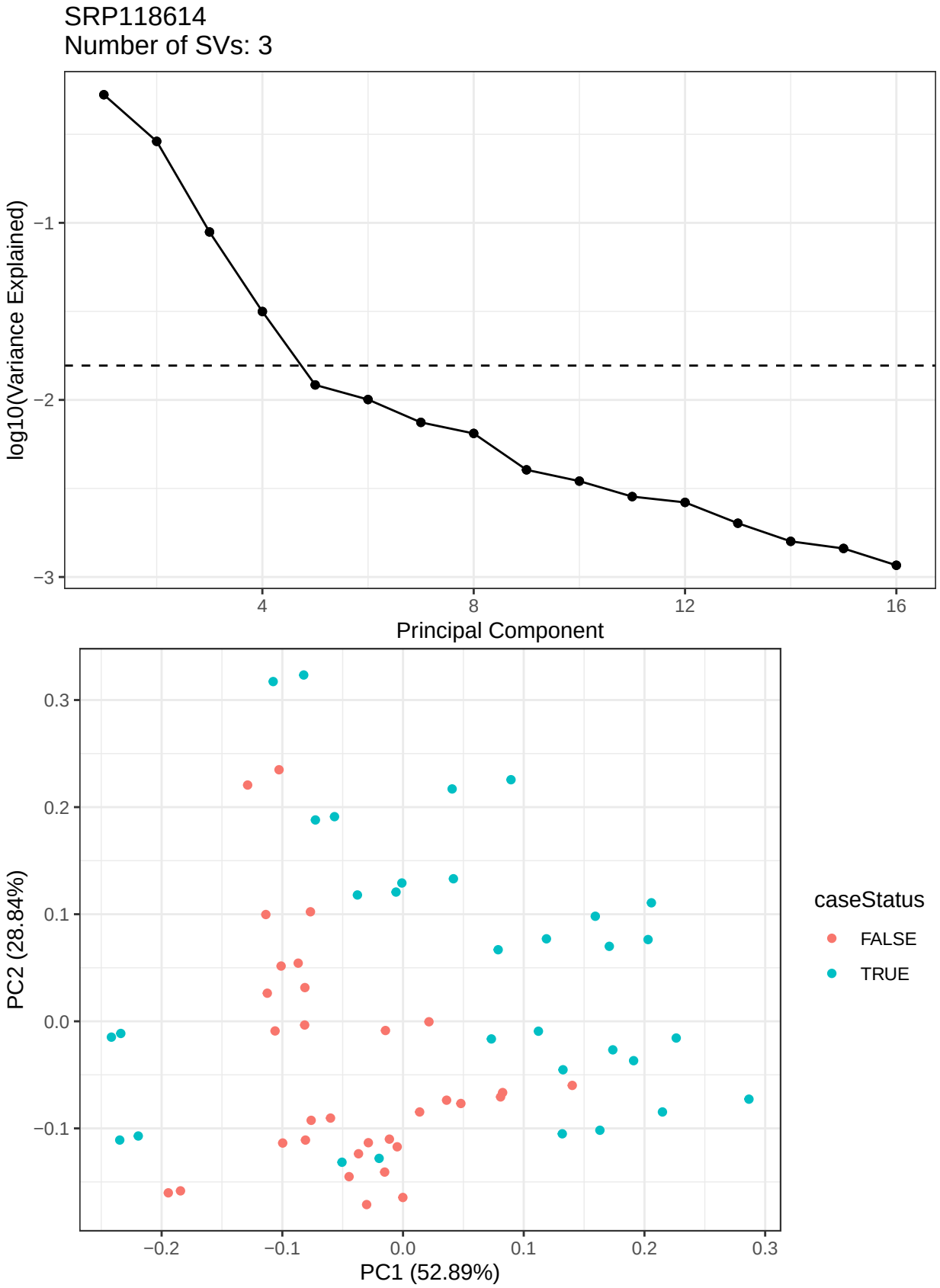
We also have a second case/control set on TB samples.

1. SRP098758 (n=428): "Samples are collected from subjects in a household contact study after a person comes back to the household, diagnosed with TB. Samples are collected every 6 months up to, 18 months. Some people go on to develop TB (cases) where as some others do not (controls). Here we are trying to establish a gene signature to predict the occurrence of TB."
2. ERP115010 (n=360): "Close contacts of active TB were defined as individuals with a cumulative duration of exposure of greater than eight hours in a confined space to the index case prior to initiation of treatment. Known human immunodeficiency virus (HIV)-positive patients were excluded. At enrollment, interferon gamma release assays (IGRAs) were done using the QuantiFERON-TB Plus assay (Qiagen, Germany), and peripheral blood was collected into Tempus tubes for whole genome transcriptional profiling by RNA sequencing. Participants who progressed to active TB were identified by linkage with the national electronic TB register. Local case notes were reviewed to identify individuals who had received preventative treatment. This submission contains data from n=360 adult participants, of which n=9 progressed to TB during a median follow-up time of 1.9 years. [note: we identify case base on IGRA assay result, which does not distinguish active vs. latent TB.]

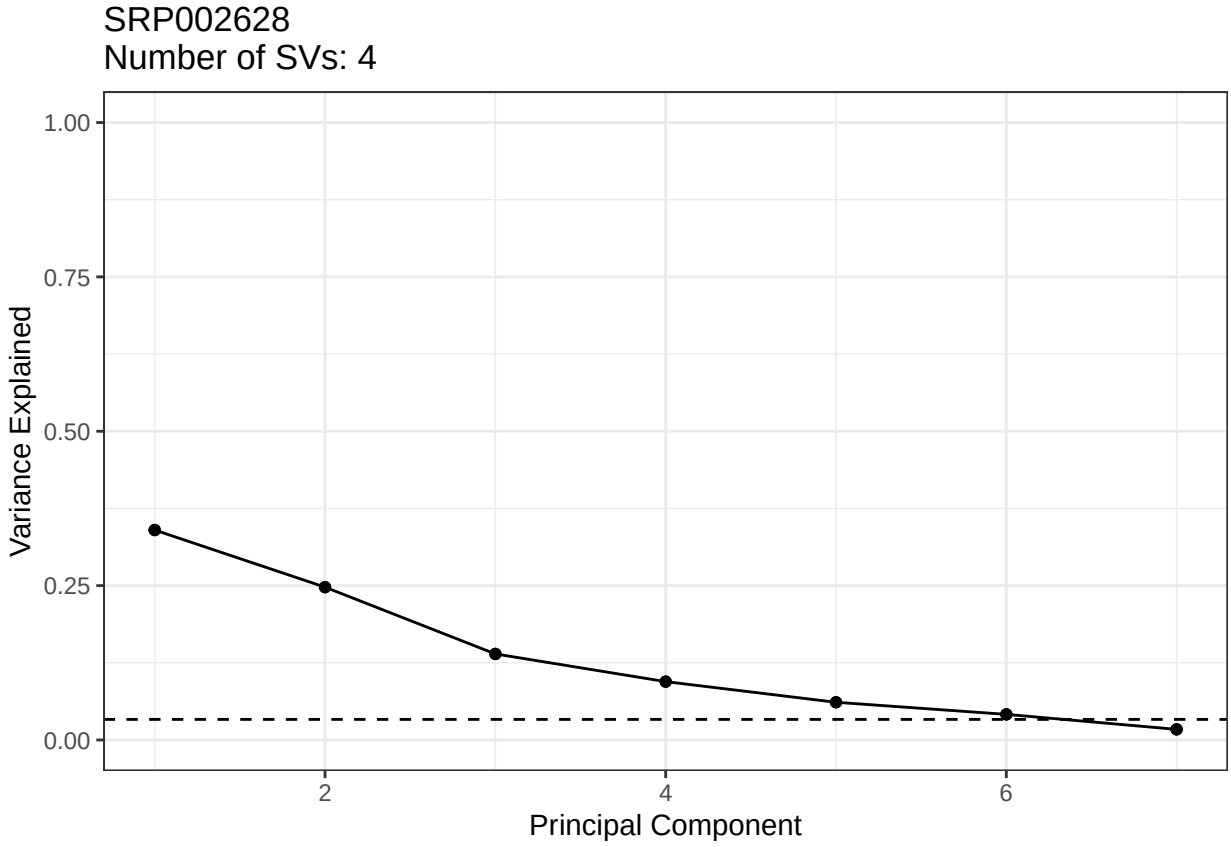
3. SRP126580 (n=54): "Overall design: We undertook RNA Sequencing (RNA-Seq) of our earlier Berry et al. 2010 (GSE19444 and GSE19442) cohorts and additionally set up a prospective cohort study at Leicester (UK) in subject groups of incident TB and recent TB contacts, respectively. In the Leicester cohort, we performed systematic longitudinal sampling and clinical characterisation first, to validate our TB signature using RNA-Seq in a new and independent cohort of individuals with active TB and LTBI, and secondly to provide longitudinal data in a low TB incidence setting. All samples in this series were re-analyzed from GSE19444. There are links on each sample page to the original sample." [*QC note: 28% of genes have count of 0*]

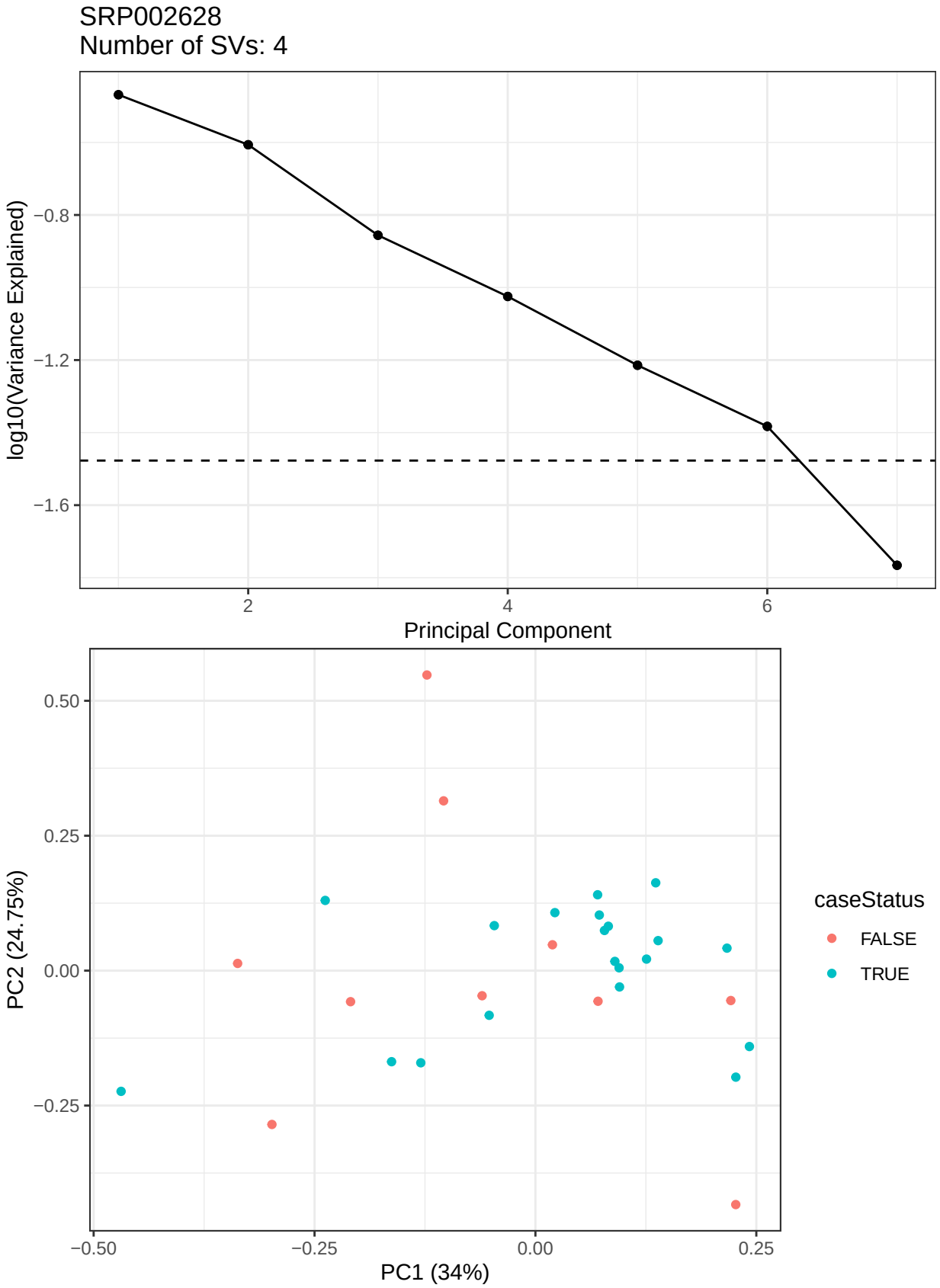
```
## 1
## [1] "SizeFactor distribution:"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.4696  0.7496  1.0697  1.1054  1.4532  2.0903
```



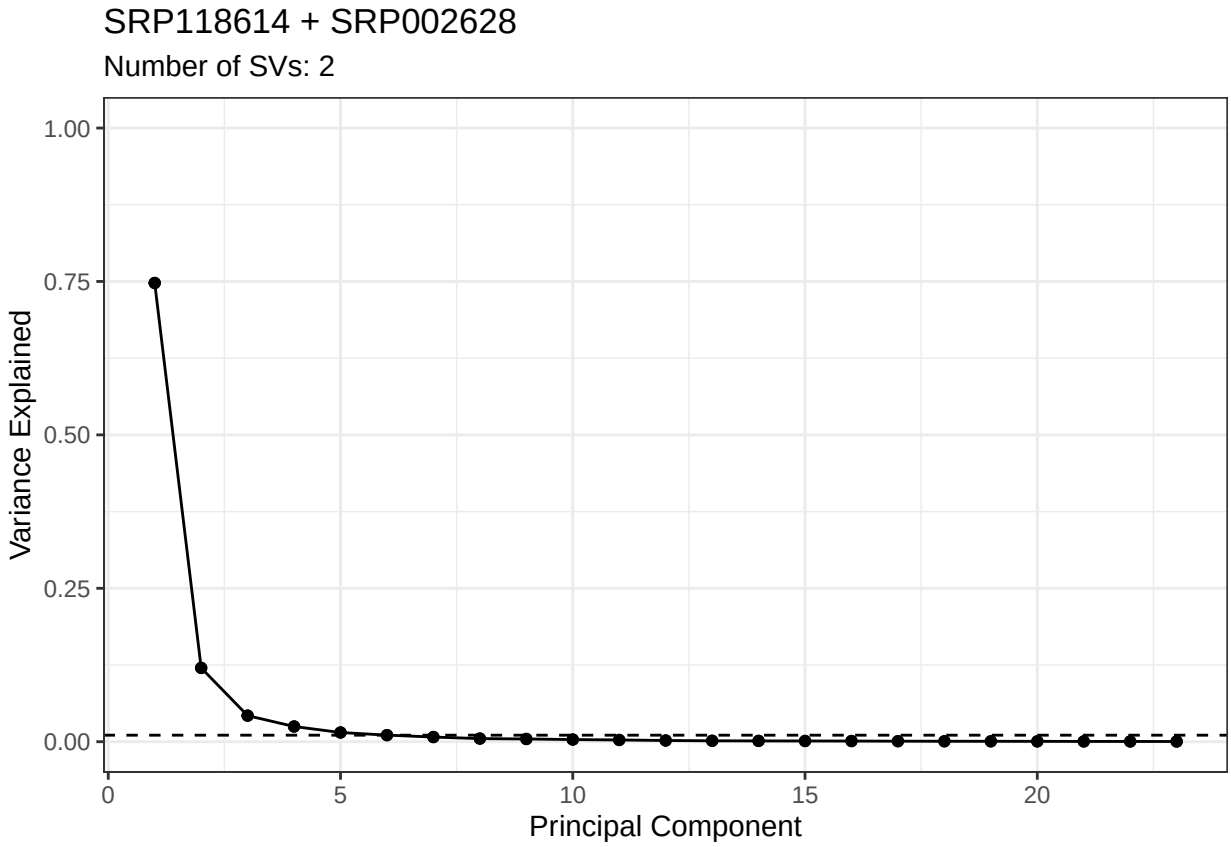

SRP118614
Number of SVs: 3
log10(Variance Explained)
−1
−2
−3
4
8
12
16
Principal Component
PC2 (28.84%)
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.1
0.0
0.1
0.2
0.3
PC1 (52.89%)
caseStatus
FALSE
TRUE

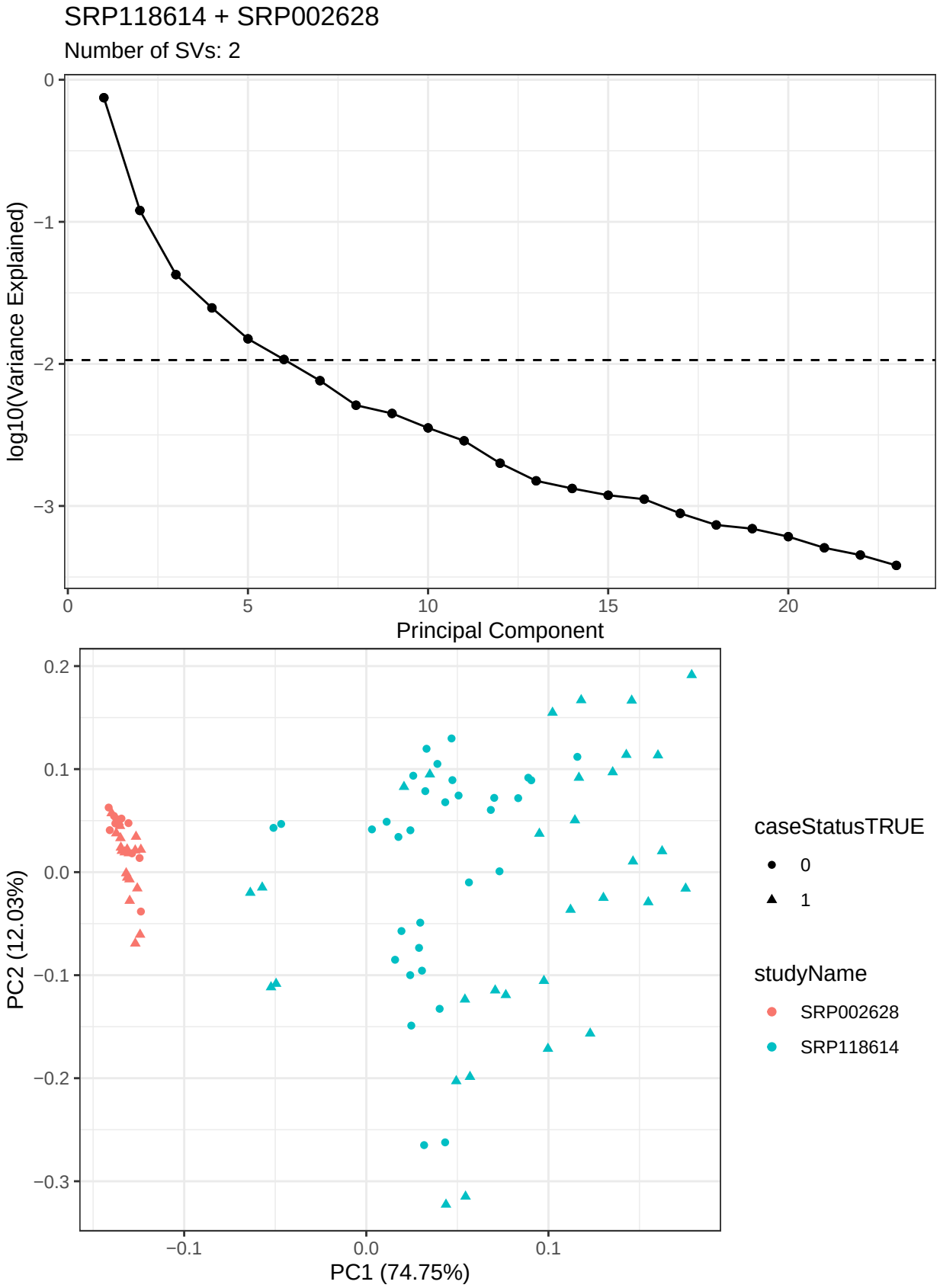
```
## Number of significant surrogate variables is:  3
## Iteration (out of 5 ):1  2  3  4  5  2
## [1] "SizeFactor distribution:"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2691  0.4948  1.4120  1.2100  1.8021  1.9976
```



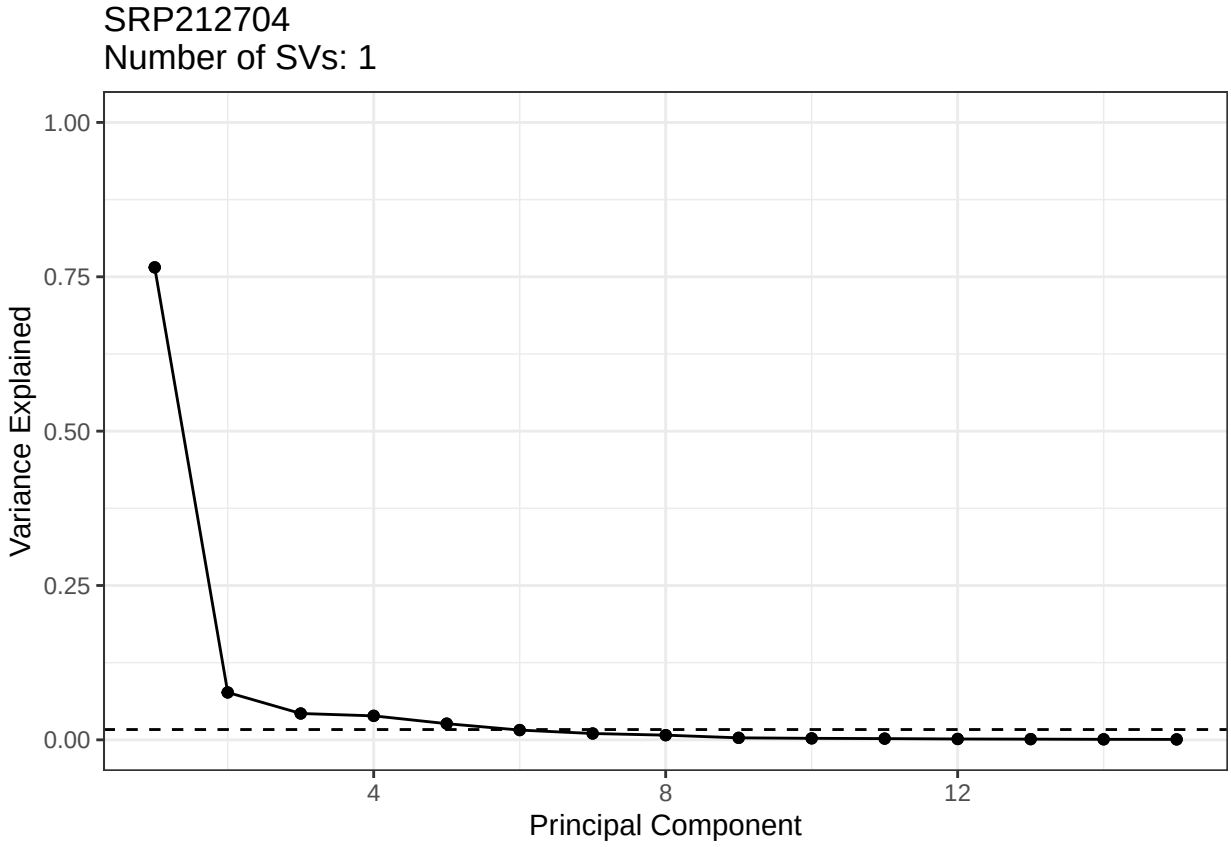

SRP002628
Number of SVs: 4
log10(Variance Explained)
−0.8
−1.2
−1.6
2
4
6
Principal Component
PC2 (24.75%)
0.50
0.25
0.00
−0.25
−0.50
−0.25
0.00
0.25
PC1 (34%)
caseStatus
FALSE
TRUE

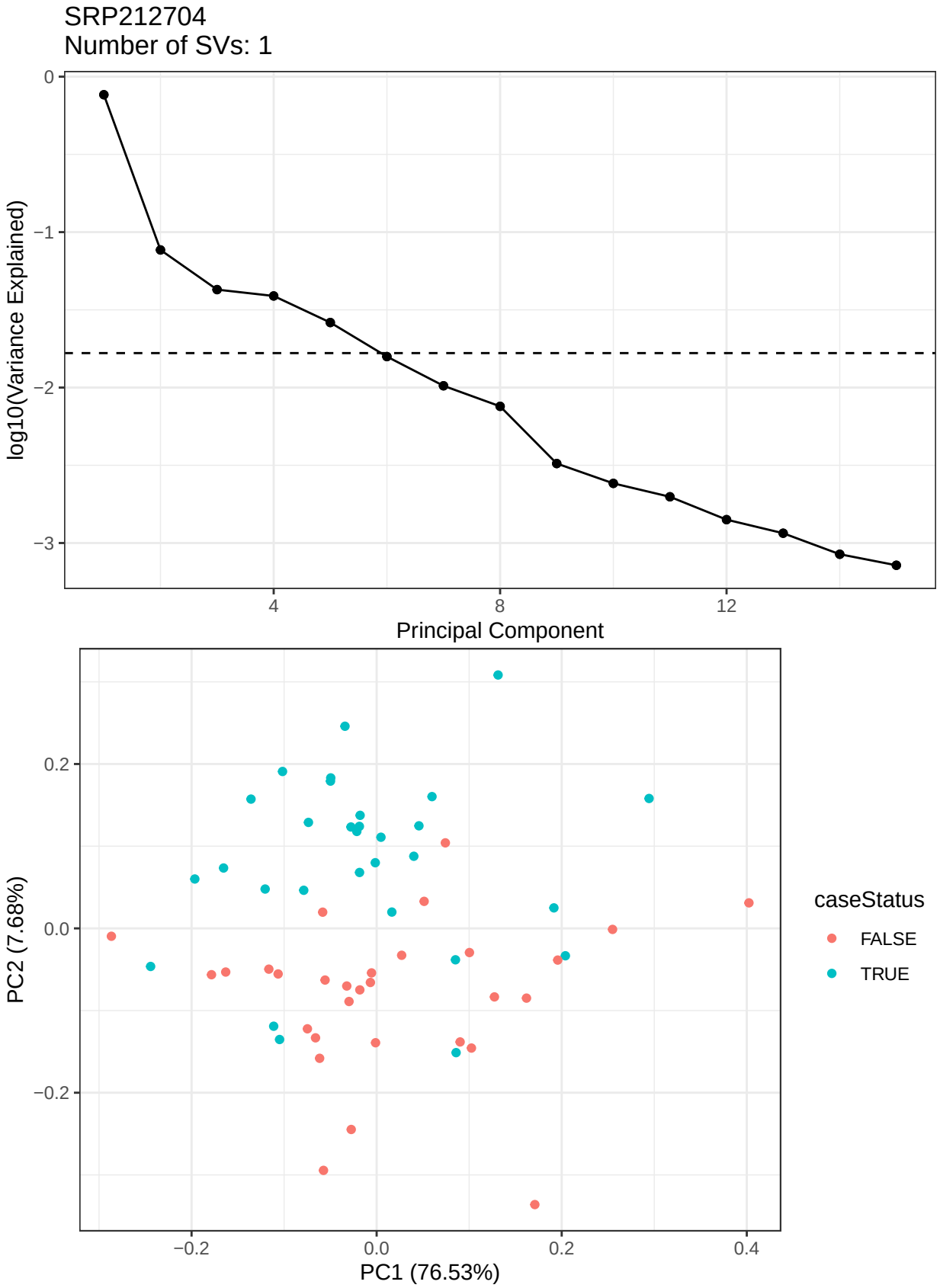
```
## Number of significant surrogate variables is:  4
## Iteration (out of 5 ):1  2  3  4  5
```



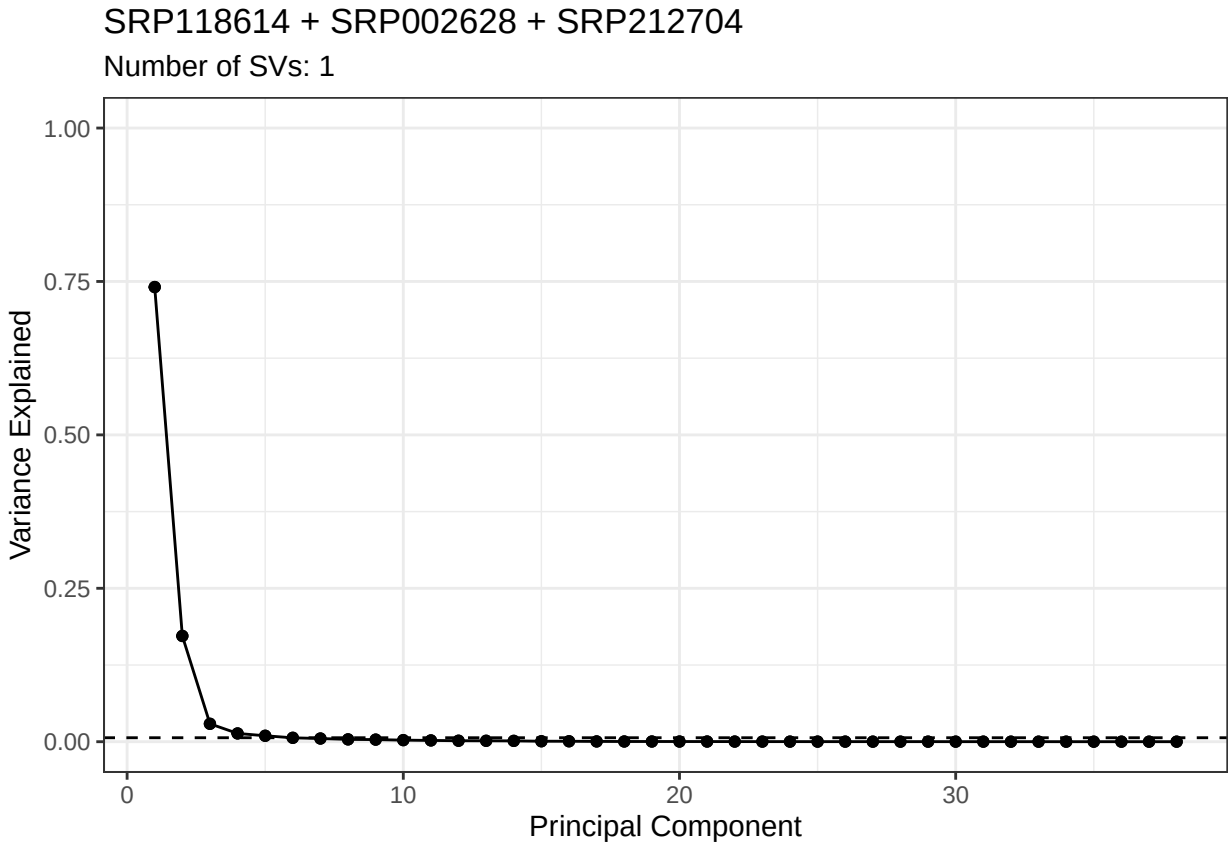

SRP118614 + SRP002628
Number of SVs: 2
log10(Variance Explained)
0
−1
−2
−3
0
5
10
15
20
Principal Component
PC2 (12.03%)
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.1
0.0
0.1
PC1 (74.75%)
caseStatusTRUE
0
1
studyName
SRP002628
SRP118614

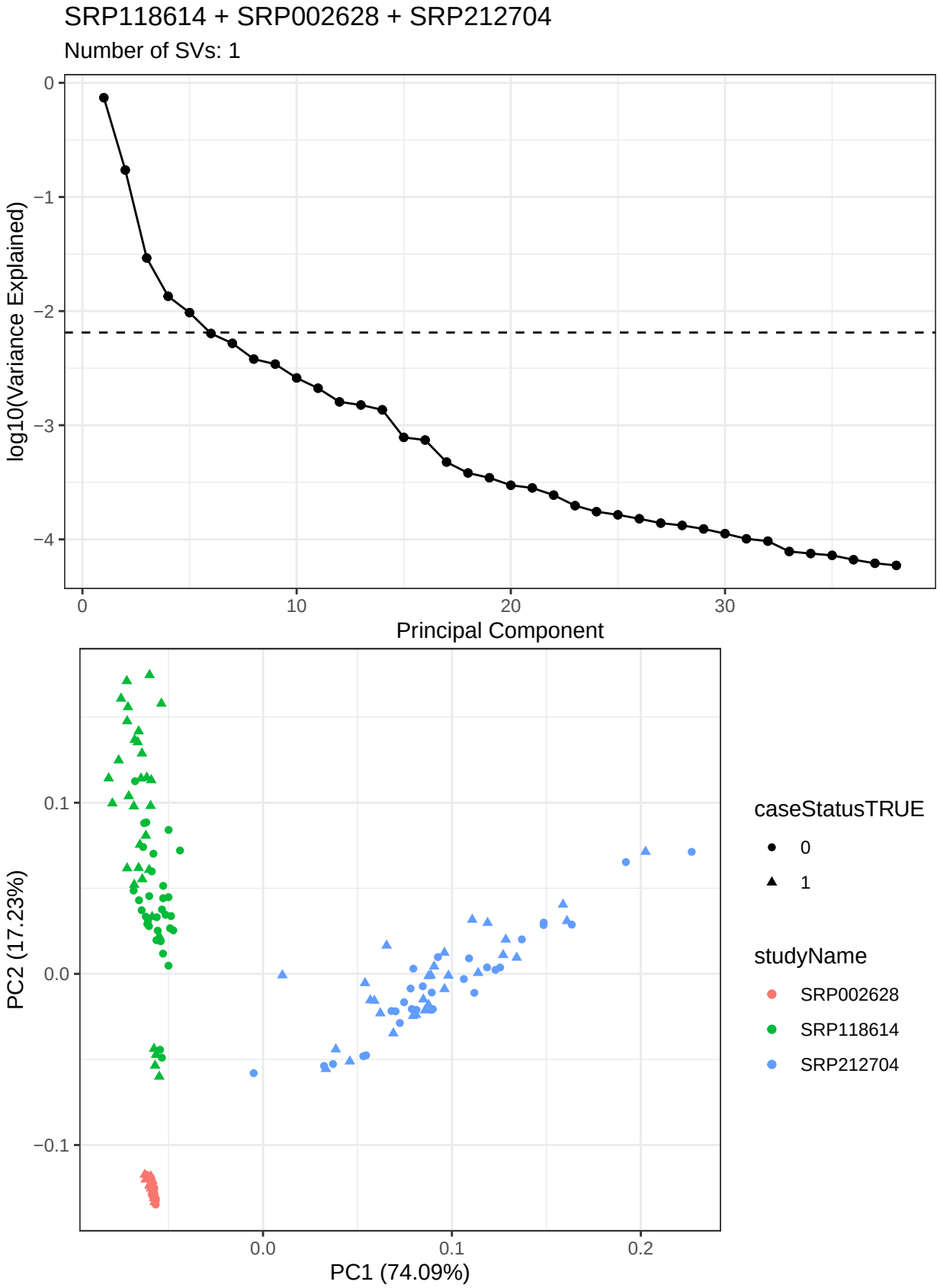
```
## Number of significant surrogate variables is:  2
## Iteration (out of 5 ):1  2  3  4  5  Number of significant surrogate variables is:  8
## Iteration (out of 5 ):1  2  3  4  5  3
## [1] "SizeFactor distribution:"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.5444  0.8273  1.0219  1.0575  1.2505  1.8502
```



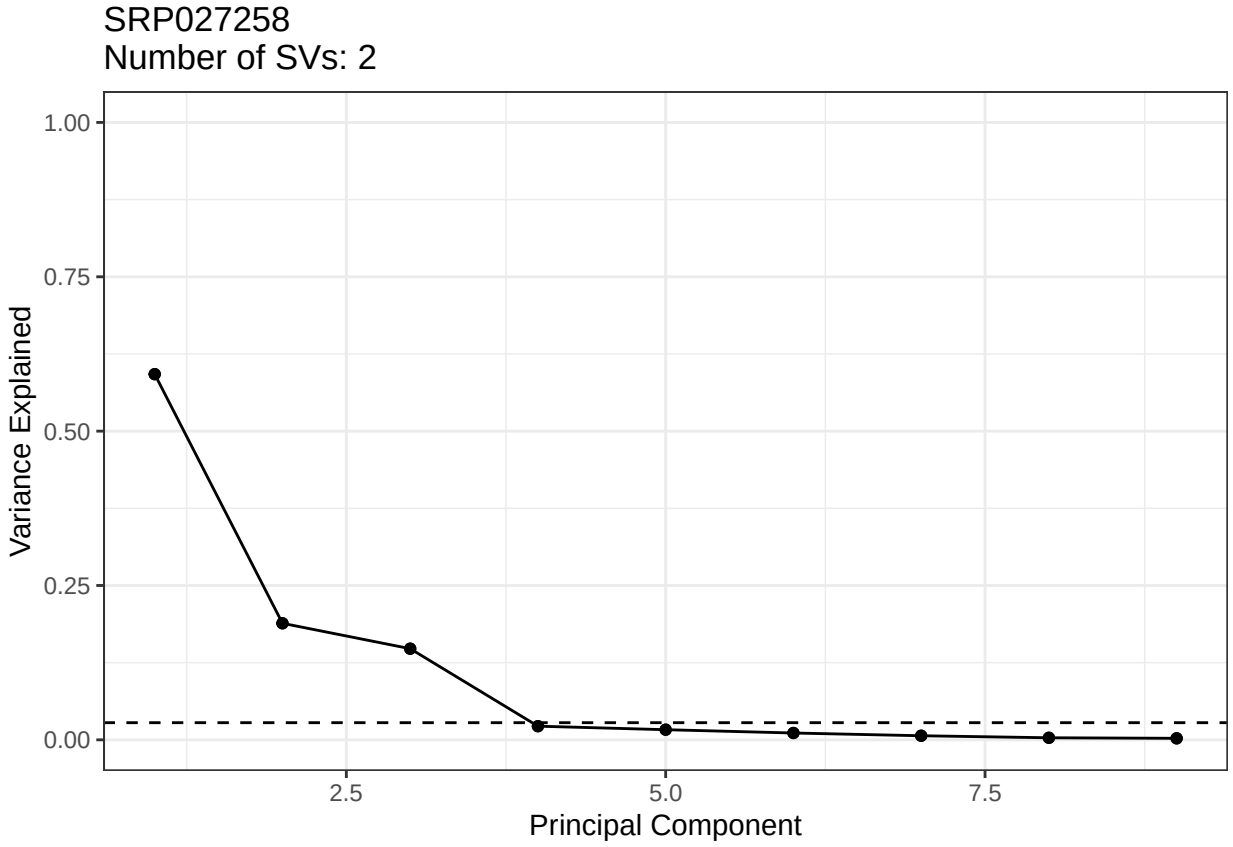

SRP212704
Number of SVs: 1
log10(Variance Explained)
0
−1
−2
−3
4
8
12
Principal Component
PC2 (7.68%)
0.2
0.0
−0.2
−0.2
0.0
0.2
0.4
PC1 (76.53%)
caseStatus
FALSE
TRUE

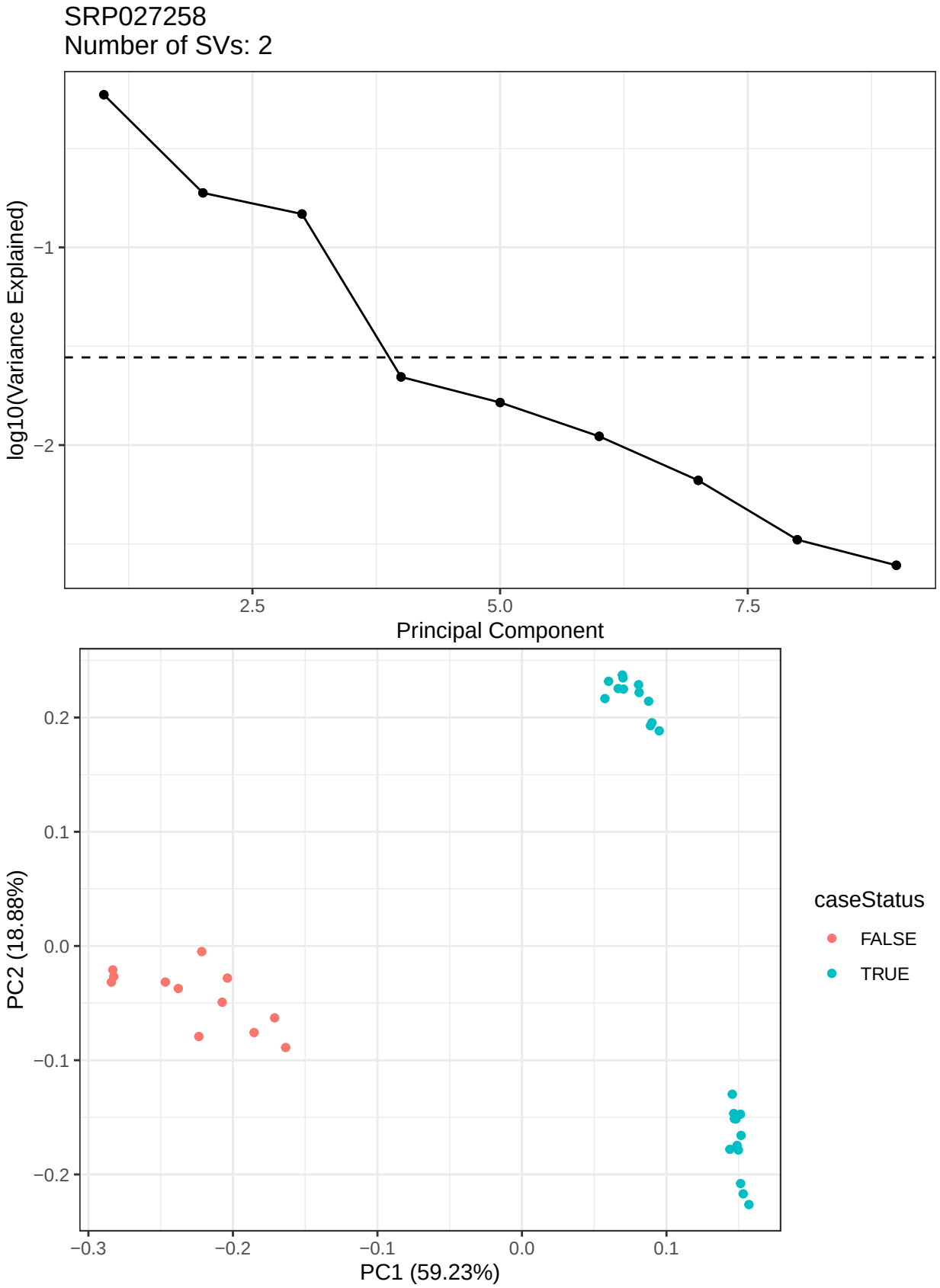
```
## Number of significant surrogate variables is:  1
## Iteration (out of 5 ):1  2  3  4  5
```



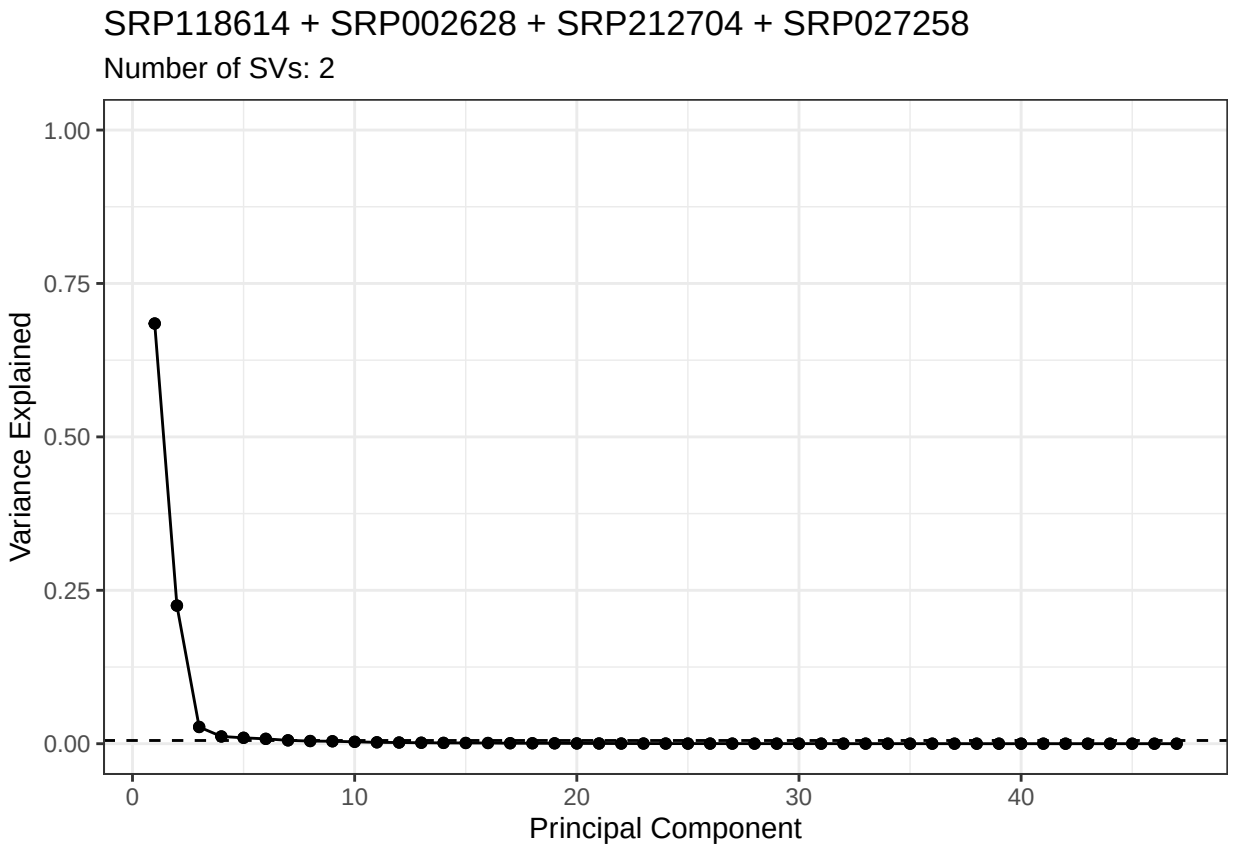

SRP118614 + SRP002628 + SRP212704
Number of SVs: 1
log10(Variance Explained)
Principal Component
PC2 (17.23%)
PC1 (74.09%)
caseStatusTRUE
0
1
studyName
SRP002628
SRP118614
SRP212704

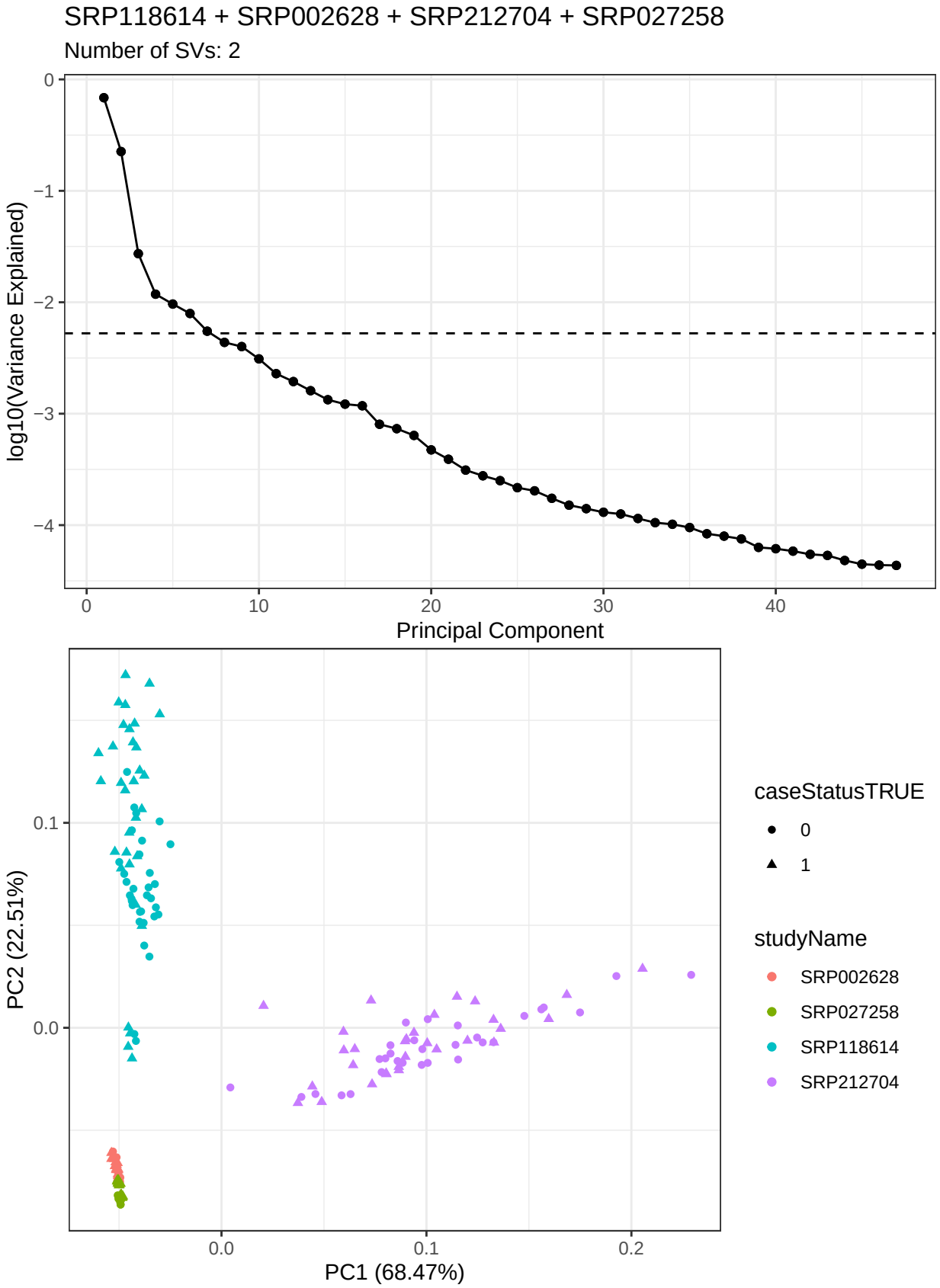
```
## Number of significant surrogate variables is:  1
## Iteration (out of 5 ):1  2  3  4  5  Number of significant surrogate variables is:  10
## Iteration (out of 5 ):1  2  3  4  5  4
## [1] "SizeFactor distribution:"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.7349  0.9231  1.0293  1.0309  1.1130  1.5355
```



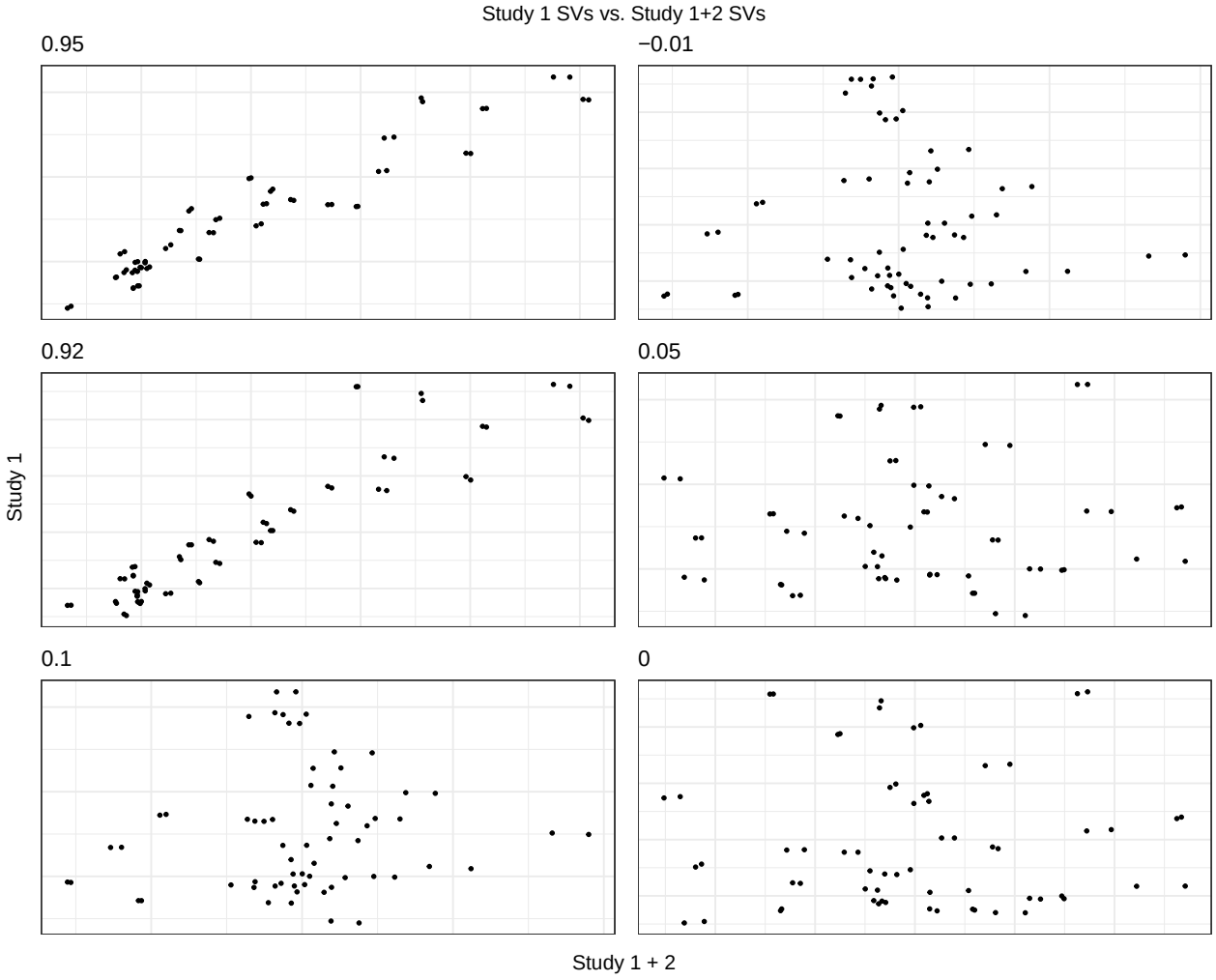

SRP027258
Number of SVs: 2
log10(Variance Explained)
−1
−2
2.5
5.0
7.5
Principal Component
PC2 (18.88%)
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.2
−0.1
0.0
0.1
PC1 (59.23%)
caseStatus
FALSE
TRUE

```
## Number of significant surrogate variables is:  2
## Iteration (out of 5 ):1  2  3  4  5
```

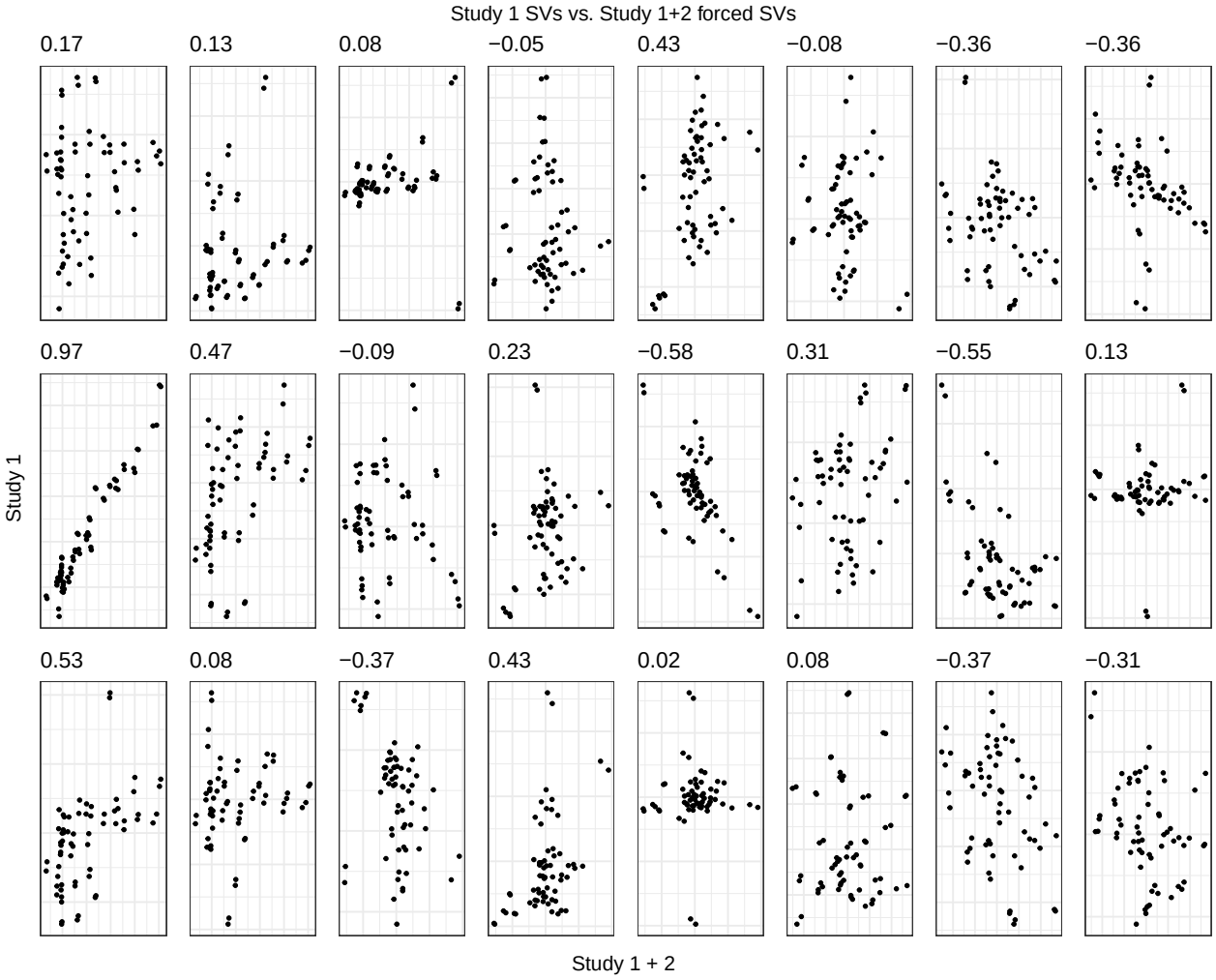



SRP118614 + SRP002628 + SRP212704 + SRP027258
Number of SVs: 2
log10(Variance Explained)
0
−1
−2
−3
−4
0
10
20
30
40
Principal Component
PC2 (22.51%)
0.1
0.0
0.0
0.1
0.2
PC1 (68.47%)
caseStatusTRUE
0
1
studyName
SRP002628
SRP027258
SRP118614
SRP212704

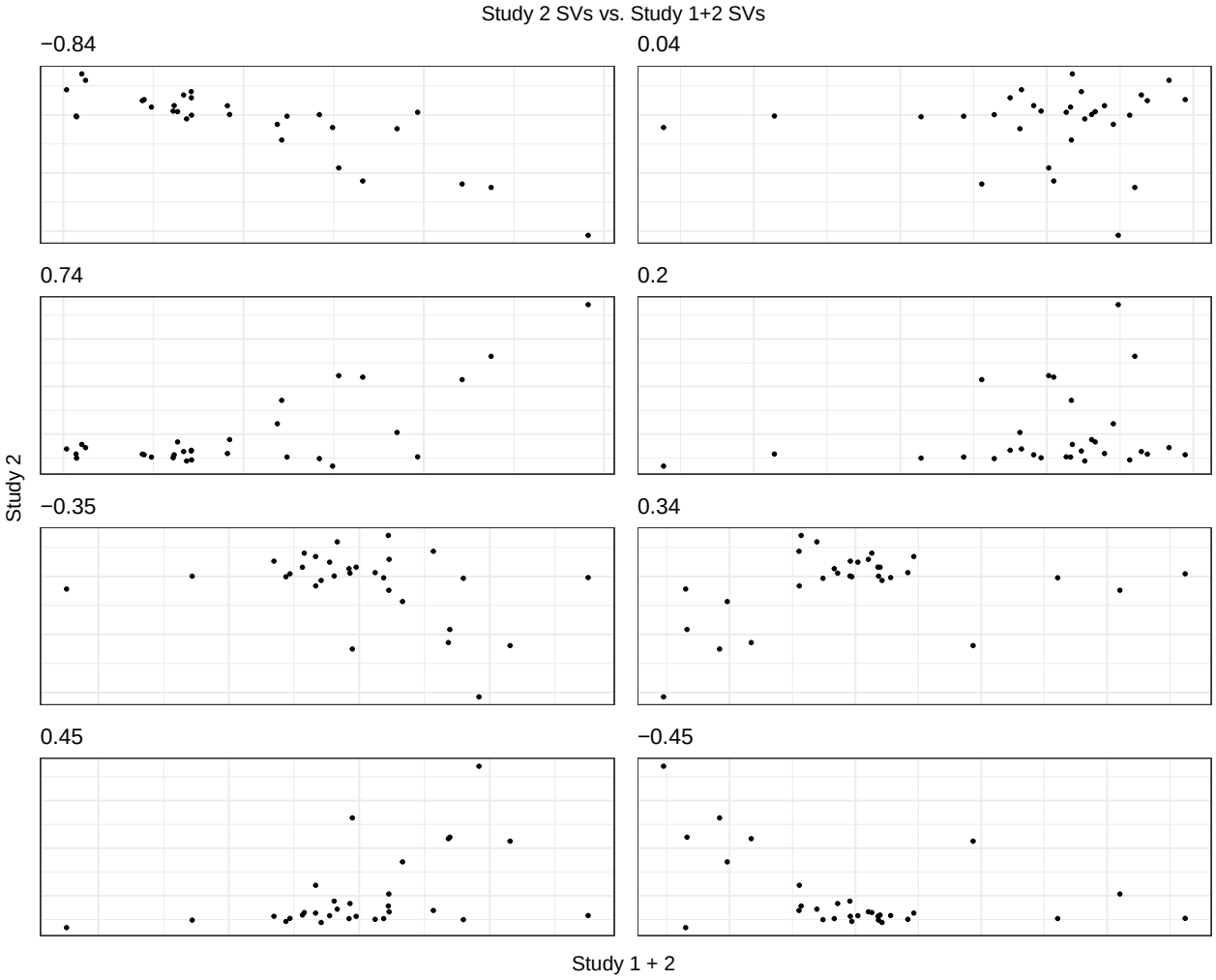
```
## Number of significant surrogate variables is:  2
## Iteration (out of 5 ):1  2  3  4  5  Number of significant surrogate variables is:  13
## Iteration (out of 5 ):1  2  3  4  5
```



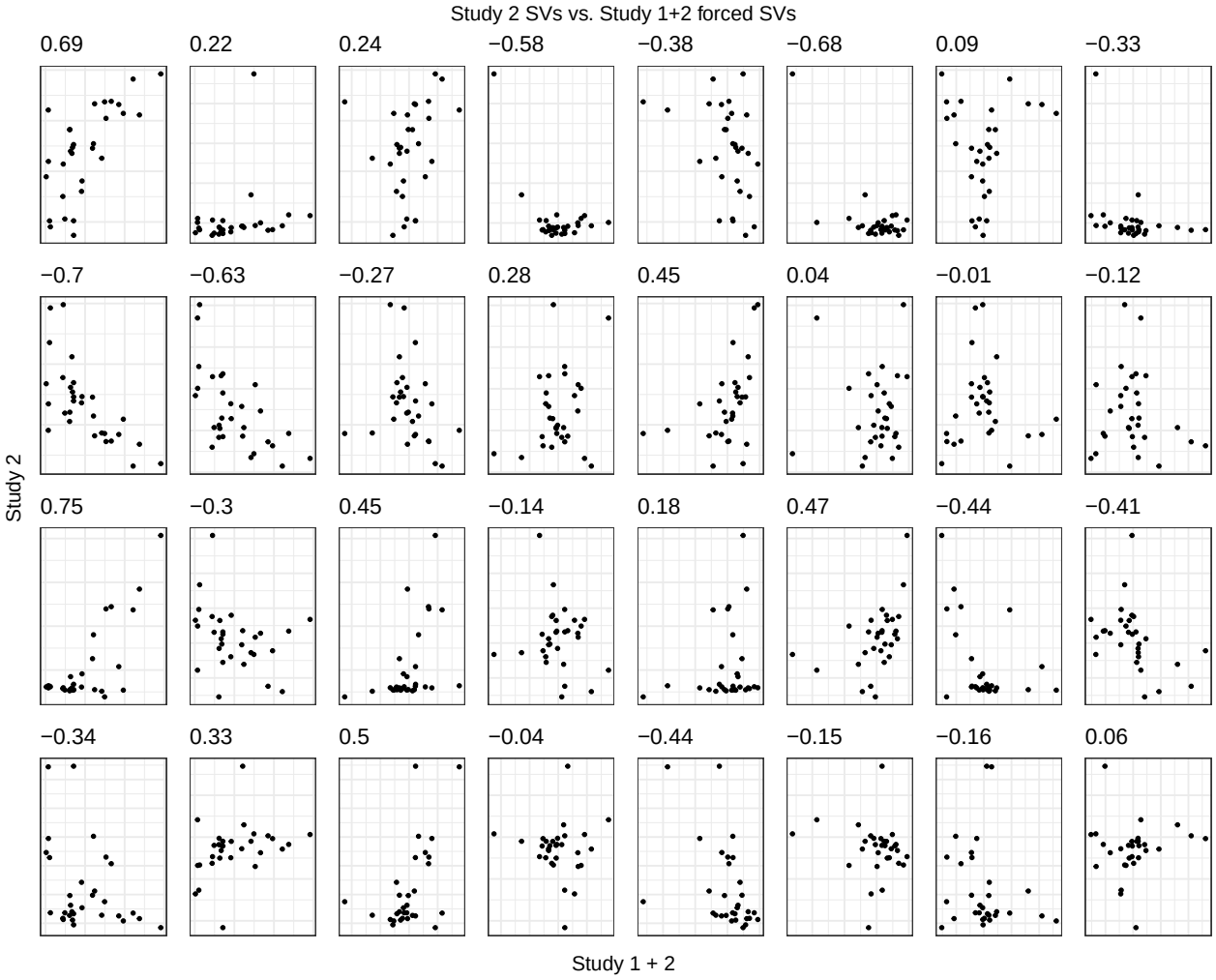

```
## [[1]]
## NULL
```


Study 1 SVs vs. Study 1+2 forced SVs
0.17
0.13
0.08
−0.05
0.43
−0.08
−0.36
−0.36
0.97
0.47
−0.09
0.23
−0.58
0.31
−0.55
0.13
0.53
0.08
−0.37
0.43
0.02
0.08
−0.37
−0.31
Study 1
Study 1 + 2

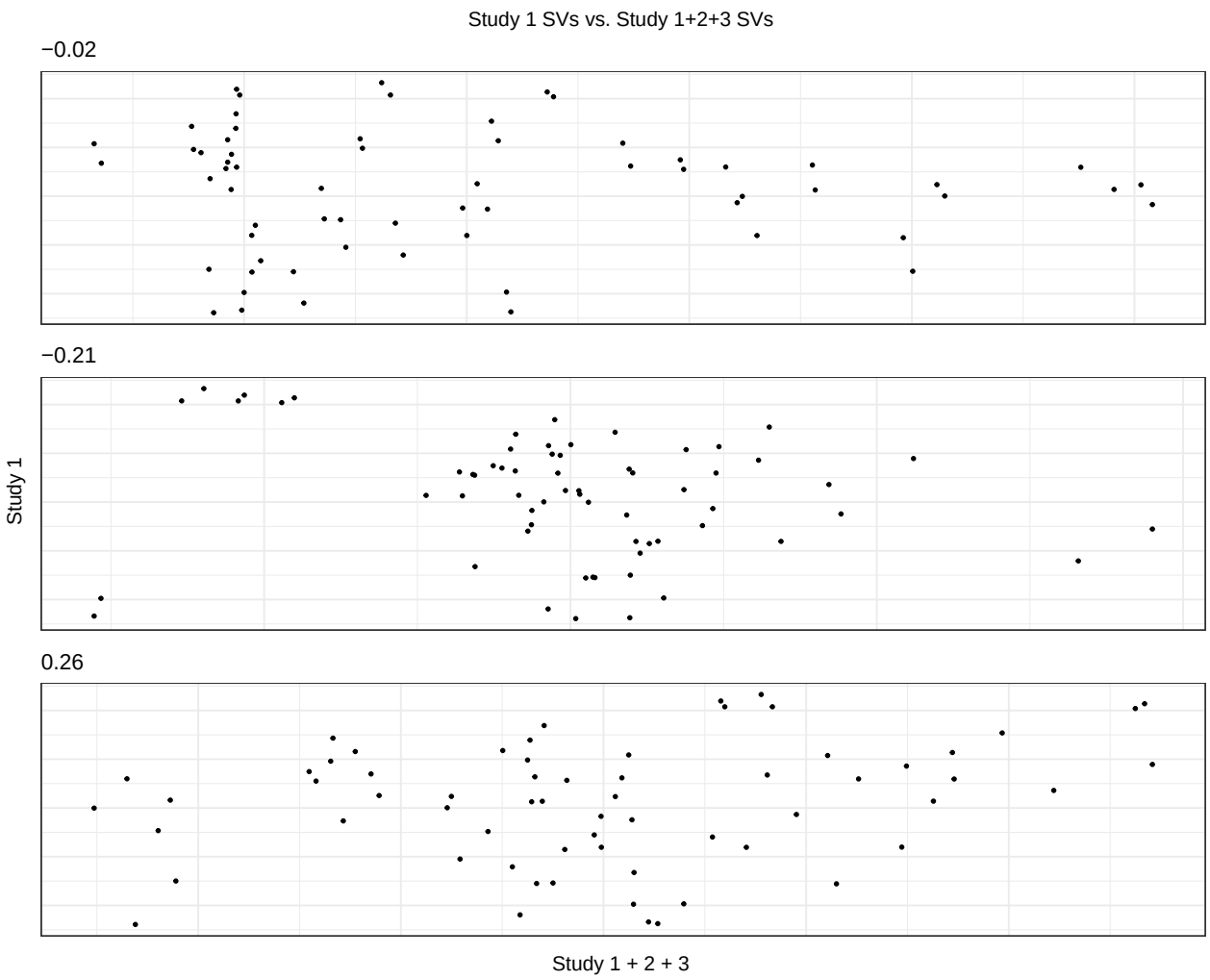
```
## [[1]]
## NULL
```


Study 2 SVs vs. Study 1+2 SVs
−0.84
0.04
0.74
0.2
−0.35
0.34
0.45
−0.45
Study 2
Study 1 + 2

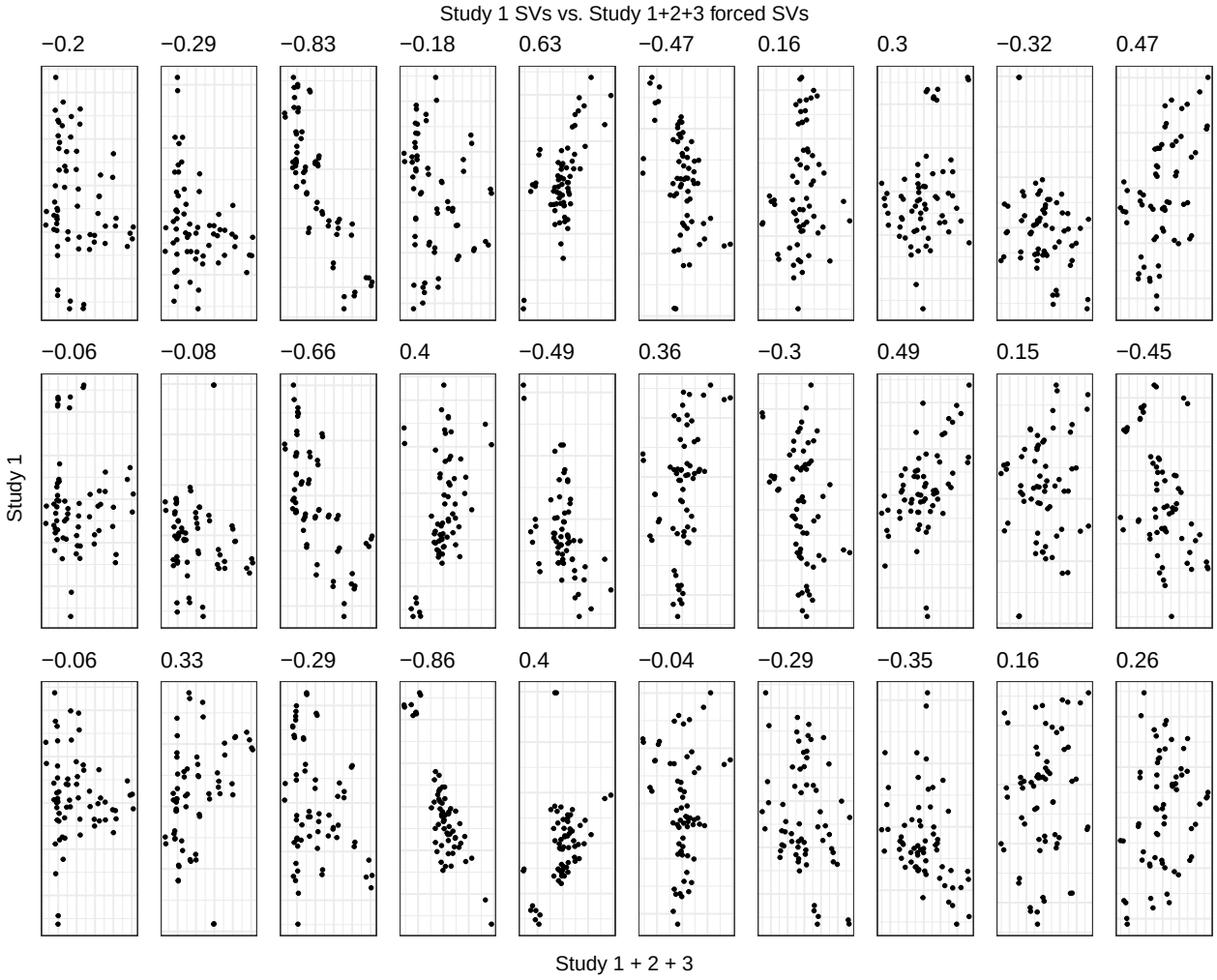
```
## [[1]]
## NULL
```


Study 2 SVs vs. Study 1+2 forced SVs
0.69
0.22
0.24
−0.58
−0.38
−0.68
0.09
−0.33
−0.7
−0.63
−0.27
0.28
0.45
0.04
−0.01
−0.12
0.75
−0.3
0.45
−0.14
0.18
0.47
−0.44
−0.41
−0.34
0.33
0.5
−0.04
−0.44
−0.15
−0.16
0.06
Study 2
Study 1 + 2

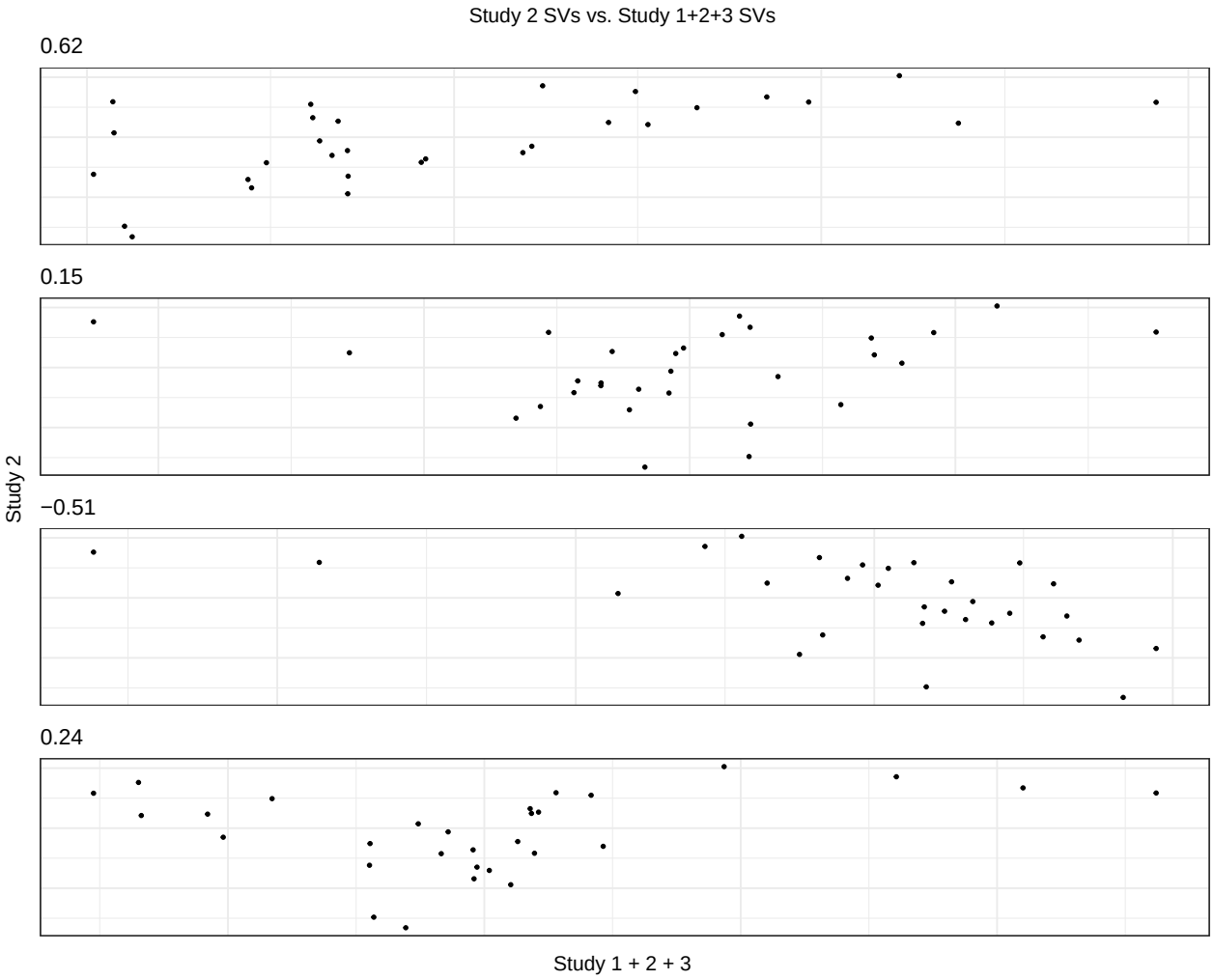
```
## [[1]]
## NULL
```



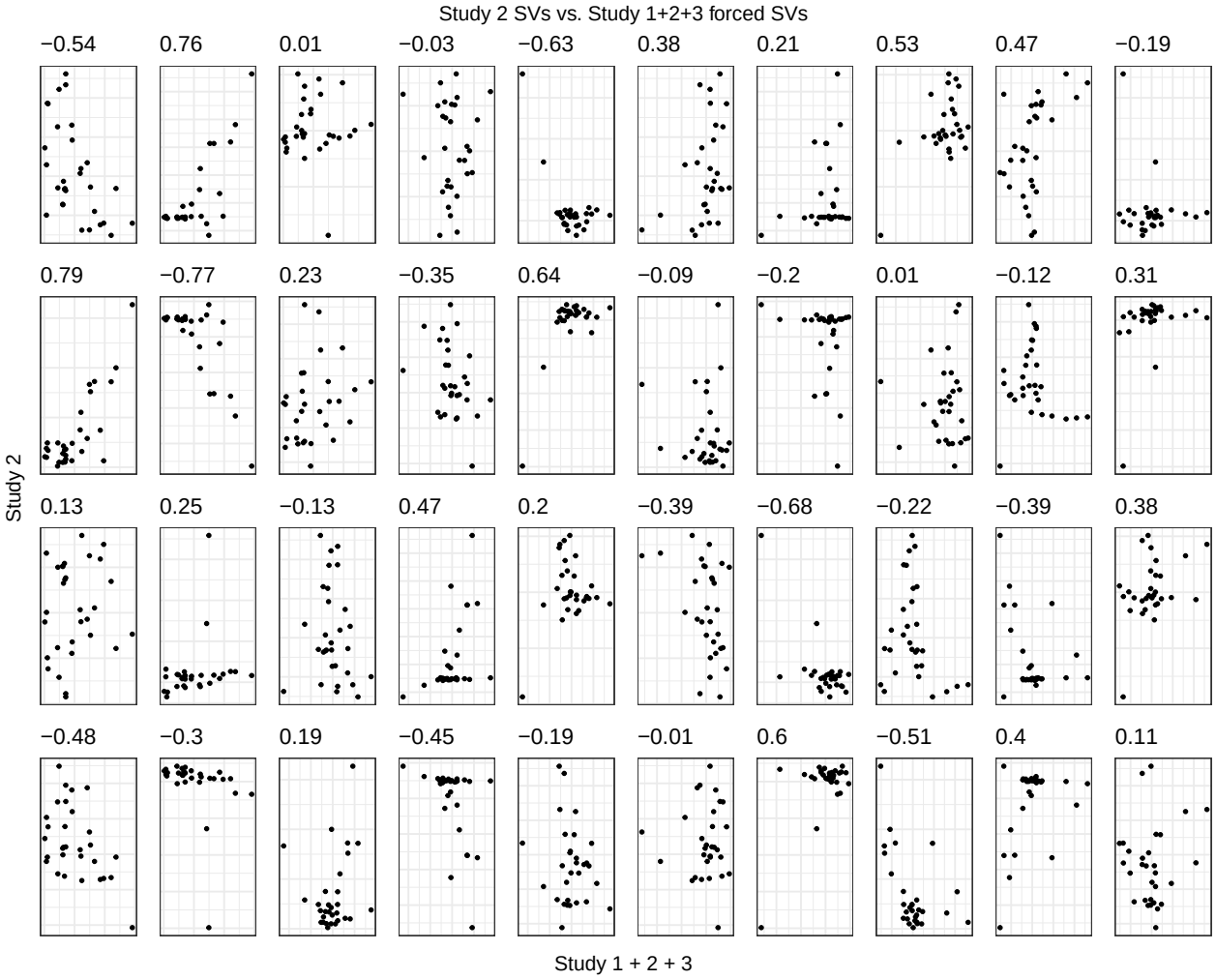

```
## [[1]]
## NULL
```


Study 1 SVs vs. Study 1+2+3 forced SVs
−0.2
−0.29
−0.83
−0.18
0.63
−0.47
0.16
0.3
−0.32
0.47
−0.06
−0.08
−0.66
0.4
−0.49
0.36
−0.3
0.49
0.15
−0.45
−0.06
0.33
−0.29
−0.86
0.4
−0.04
−0.29
−0.35
0.16
0.26
Study 1
Study 1 + 2 + 3

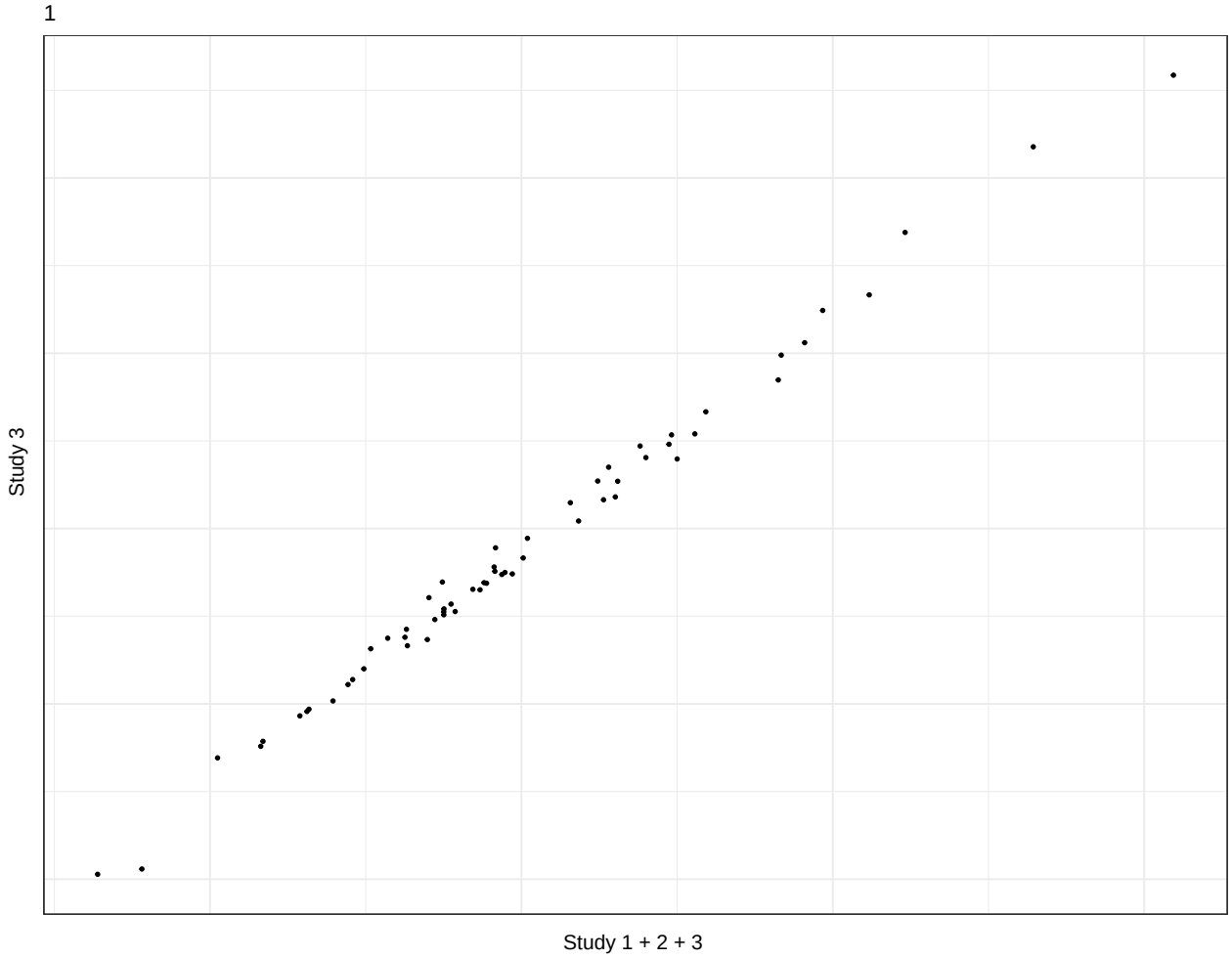
```
## [[1]]
## NULL
```



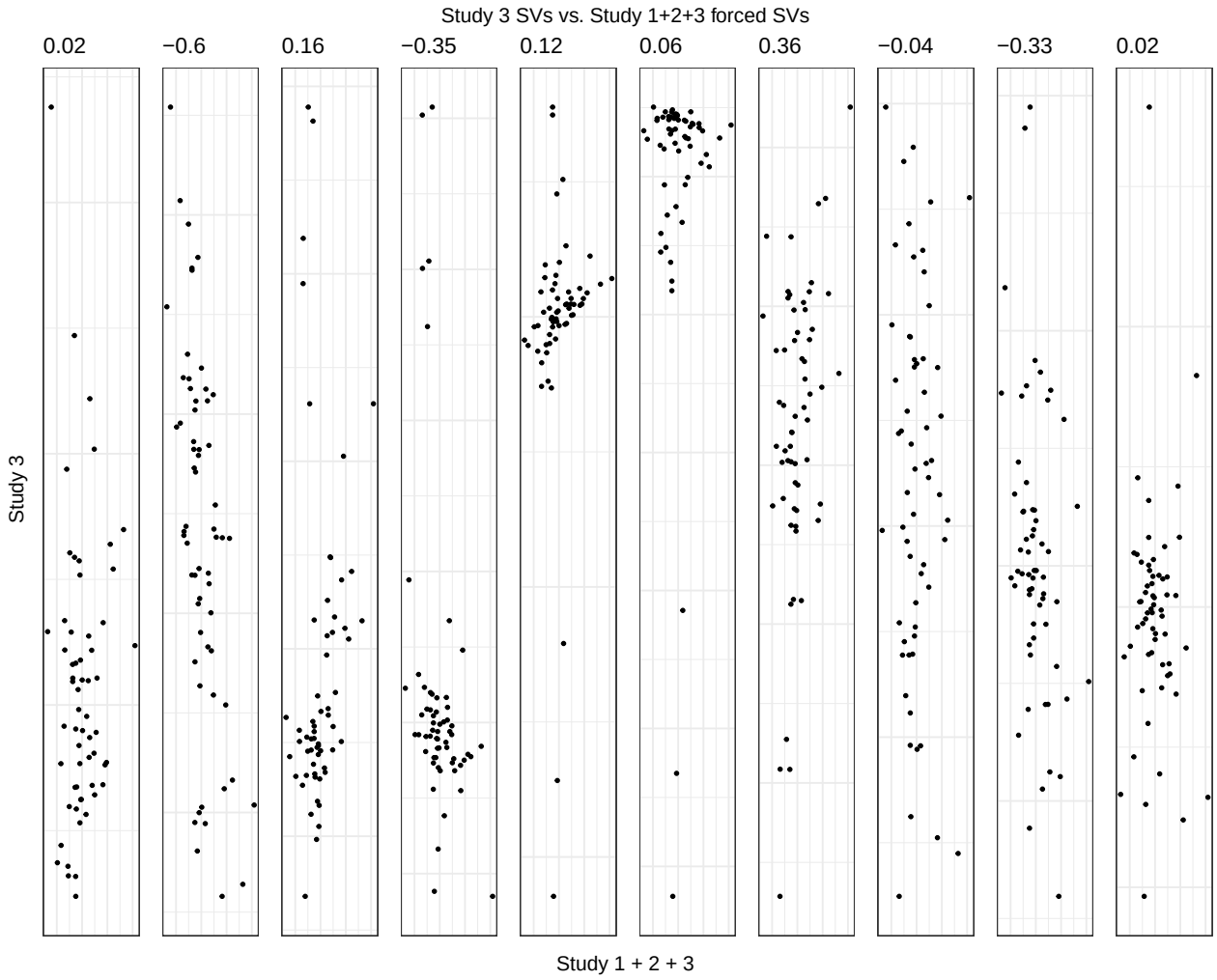

```
## [[1]]
## NULL
```


Study 2 SVs vs. Study 1+2+3 forced SVs
−0.54
0.76
0.01
−0.03
−0.63
0.38
0.21
0.53
0.47
−0.19
0.79
−0.77
0.23
−0.35
0.64
−0.09
−0.2
0.01
−0.12
0.31
0.13
0.25
−0.13
0.47
0.2
−0.39
−0.68
−0.22
−0.39
0.38
−0.48
−0.3
0.19
−0.45
−0.19
−0.01
0.6
−0.51
0.4
0.11
Study 2
Study 1 + 2 + 3

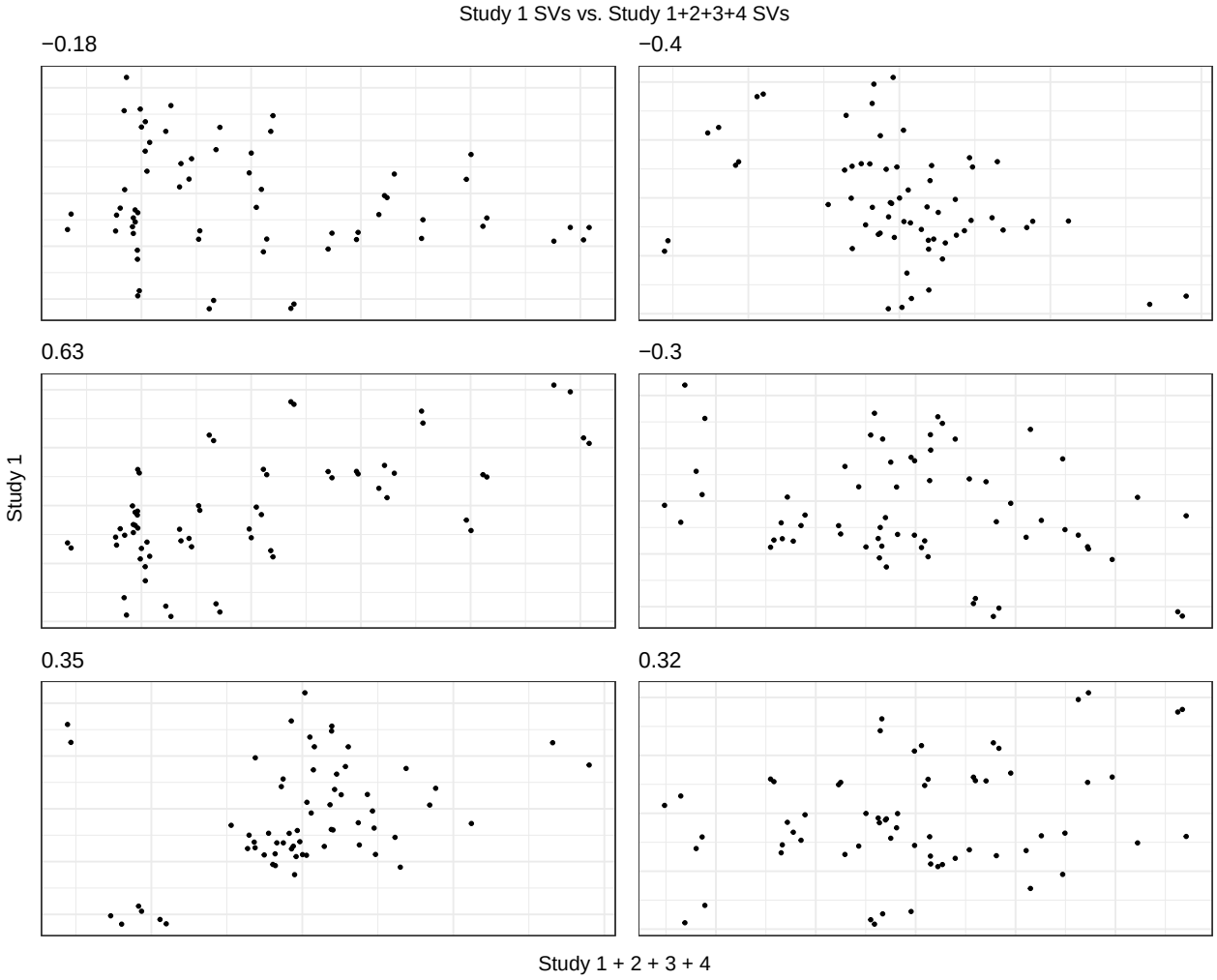
```
## [[1]]
## NULL
```



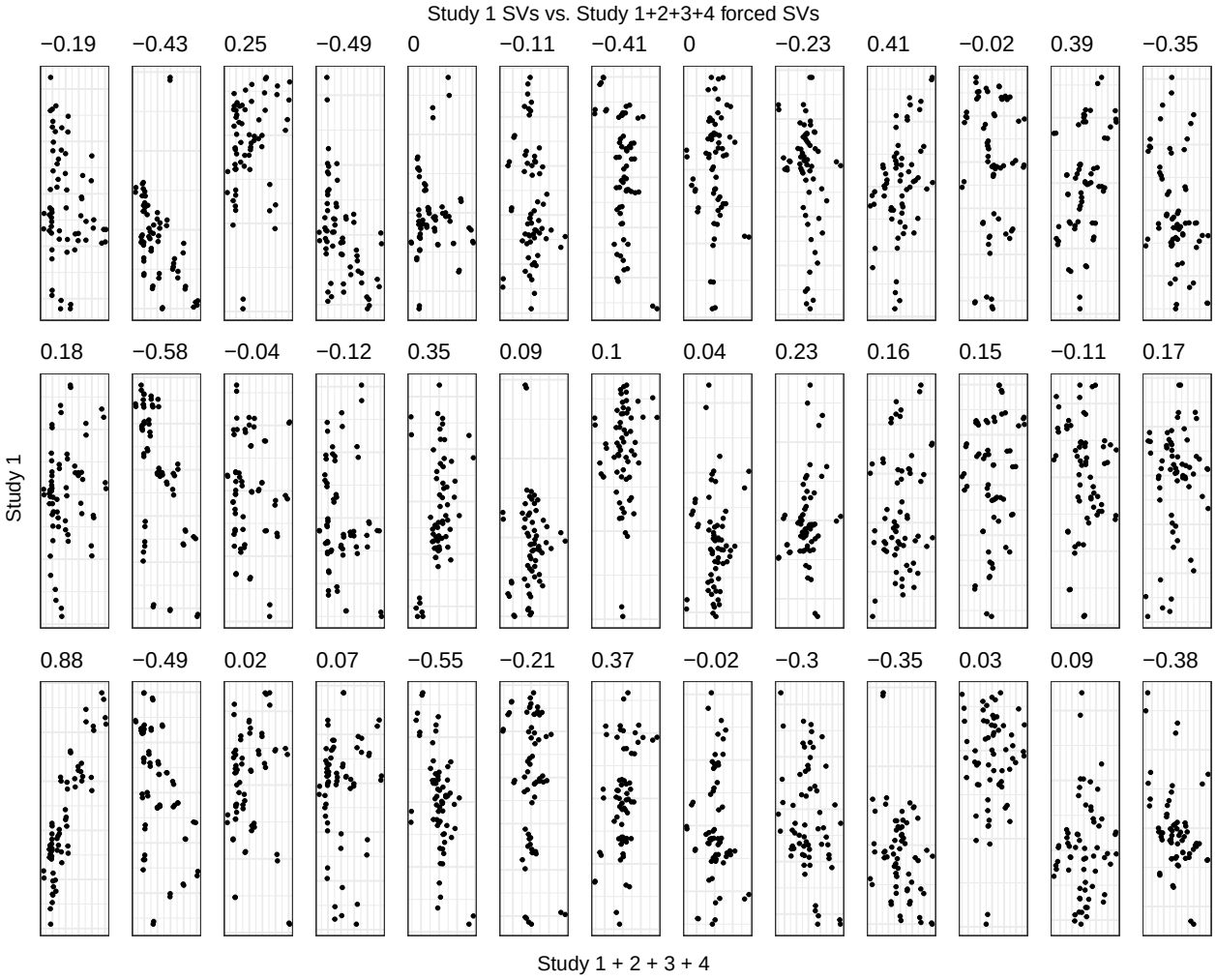

```
## [[1]]
## NULL
```


Study 3 SVs vs. Study 1+2+3 forced SVs
0.02
−0.6
0.16
−0.35
0.12
0.06
0.36
−0.04
−0.33
0.02
Study 3
Study 1 + 2 + 3

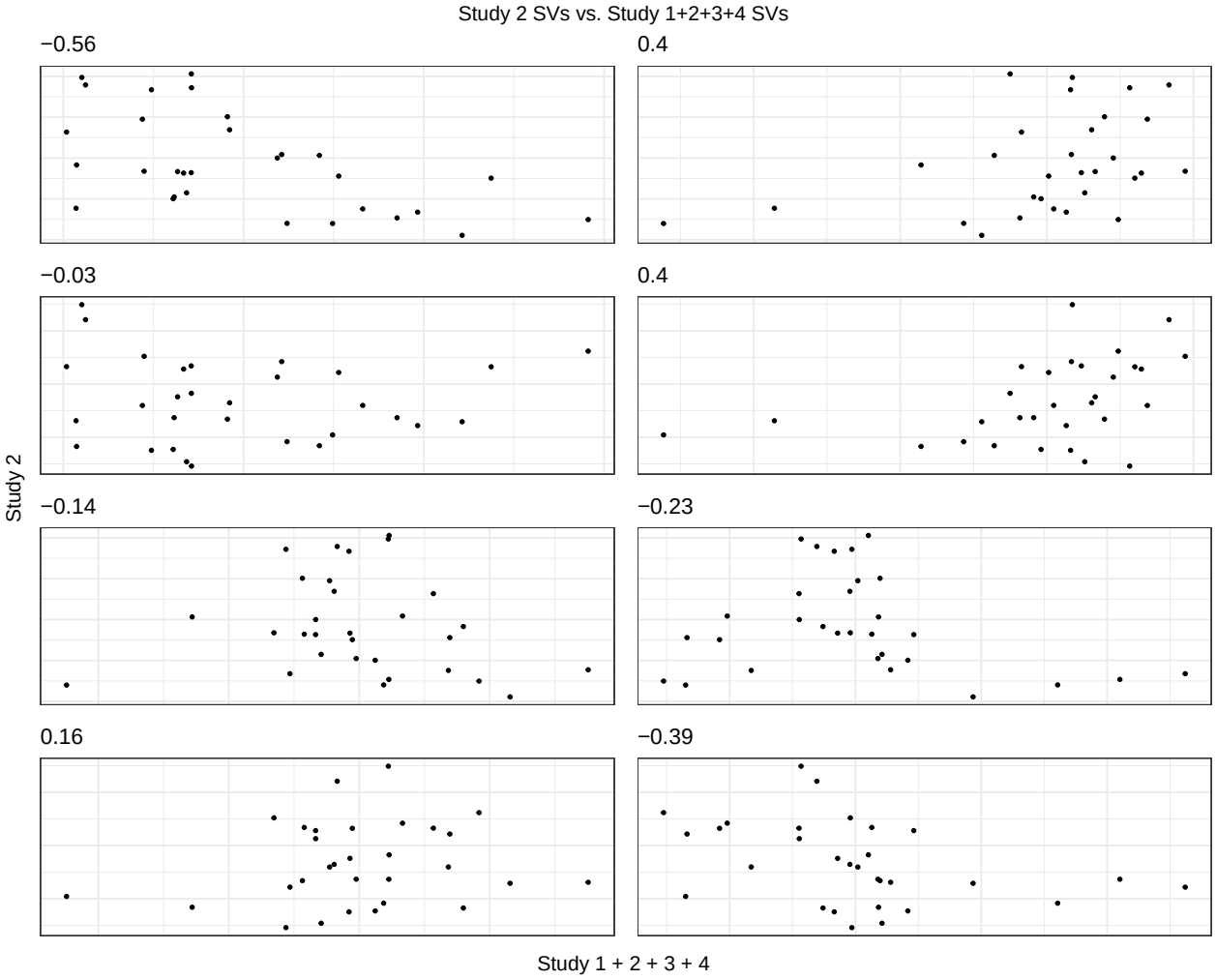
```
## [[1]]
## NULL
```


Study 1 SVs vs. Study 1+2+3+4 SVs
−0.18
−0.4
0.63
−0.3
0.35
0.32
Study 1
Study 1 + 2 + 3 + 4

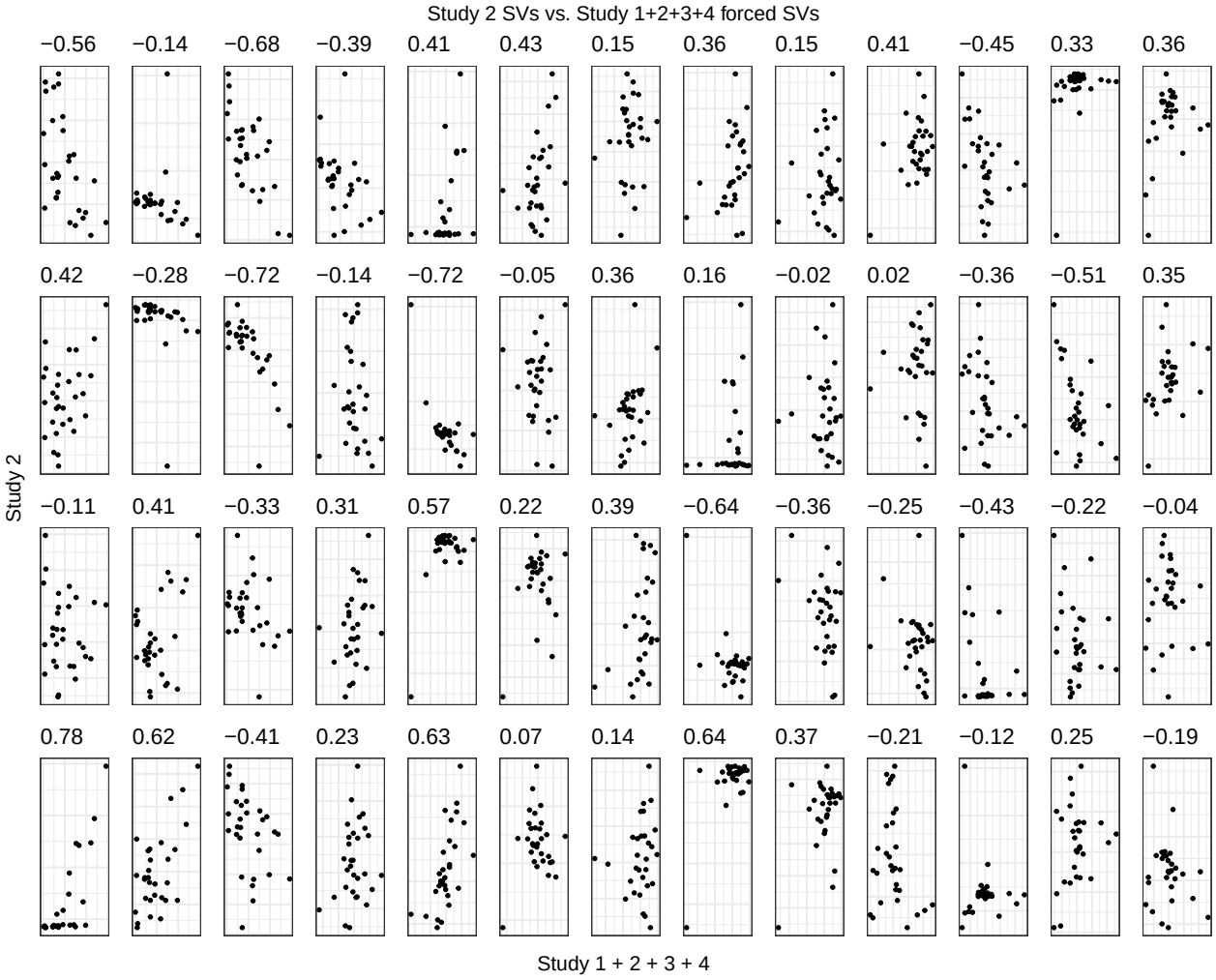
```
## [[1]]
## NULL
```


Study 1 SVs vs. Study 1+2+3+4 forced SVs
−0.19
−0.43
0.25
−0.49
0
−0.11
−0.41
0
−0.23
0.41
−0.02
0.39
−0.35
0.18
−0.58
−0.04
−0.12
0.35
0.09
0.1
0.04
0.23
0.16
0.15
−0.11
0.17
0.88
−0.49
0.02
0.07
−0.55
−0.21
0.37
−0.02
−0.3
−0.35
0.03
0.09
−0.38
Study 1
Study 1 + 2 + 3 + 4

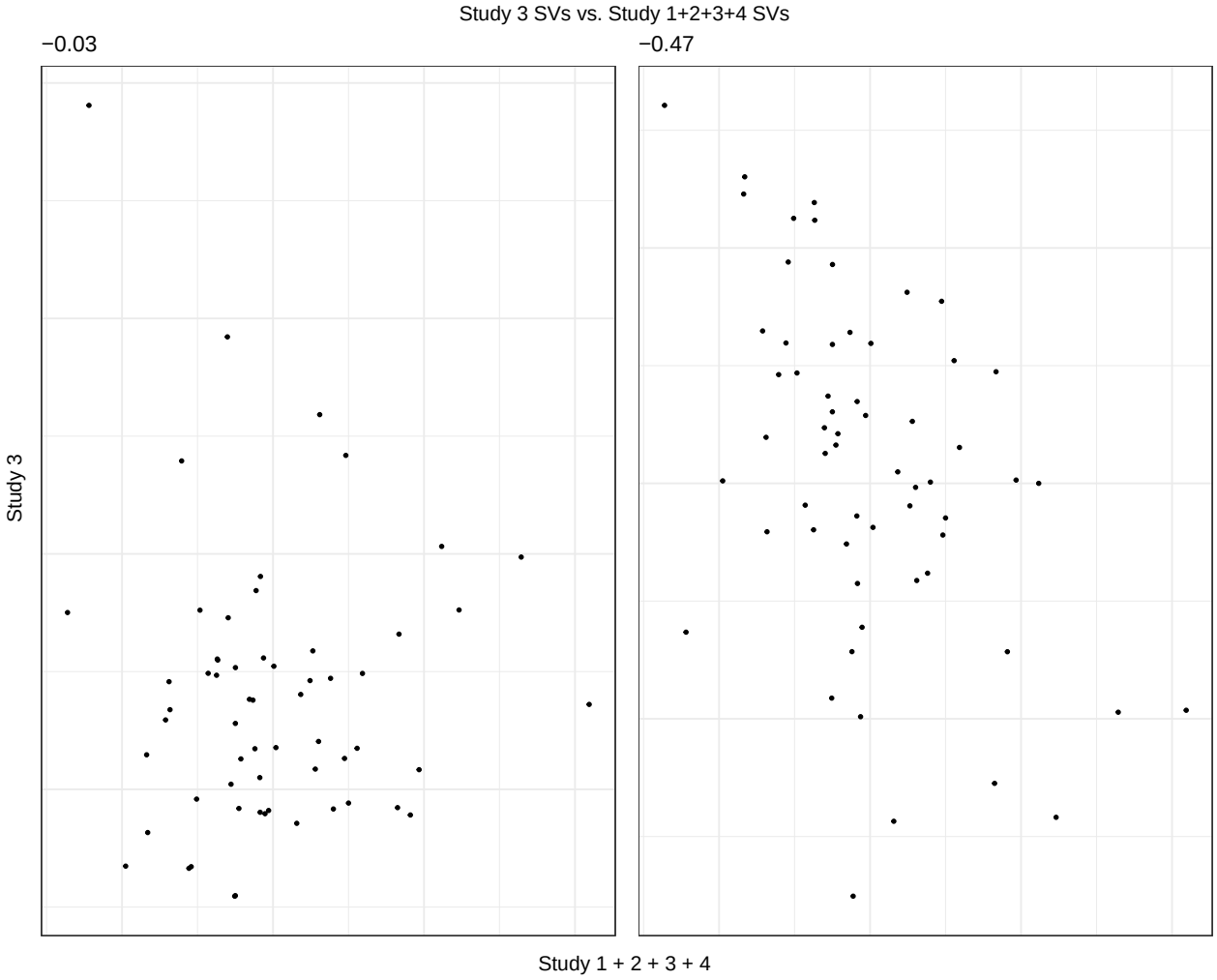
```
## [[1]]
## NULL
```


Study 2 SVs vs. Study 1+2+3+4 SVs
−0.56
0.4
−0.03
0.4
−0.14
−0.23
0.16
−0.39
Study 2
Study 1 + 2 + 3 + 4

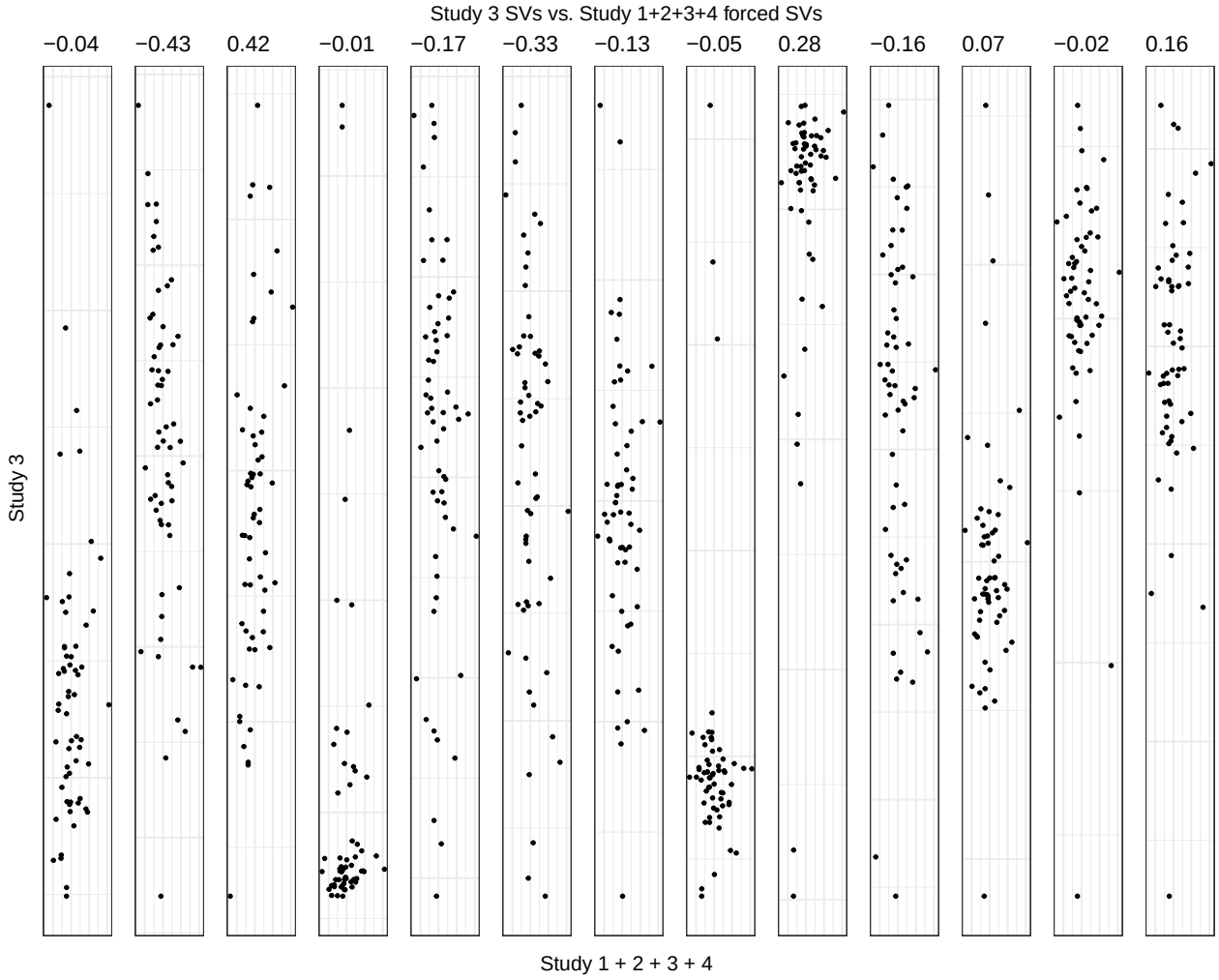
```
## [[1]]
## NULL
```


Study 2 SVs vs. Study 1+2+3+4 forced SVs
−0.56 −0.14 −0.68 −0.39 0.41 0.43 0.15 0.36 0.15 0.41 −0.45 0.33 0.36
0.42 −0.28 −0.72 −0.14 −0.72 −0.05 0.36 0.16 −0.02 0.02 −0.36 −0.51 0.35
−0.11 0.41 −0.33 0.31 0.57 0.22 0.39 −0.64 −0.36 −0.25 −0.43 −0.22 −0.04
0.78 0.62 −0.41 0.23 0.63 0.07 0.14 0.64 0.37 −0.21 −0.12 0.25 −0.19
Study 2
Study 1 + 2 + 3 + 4

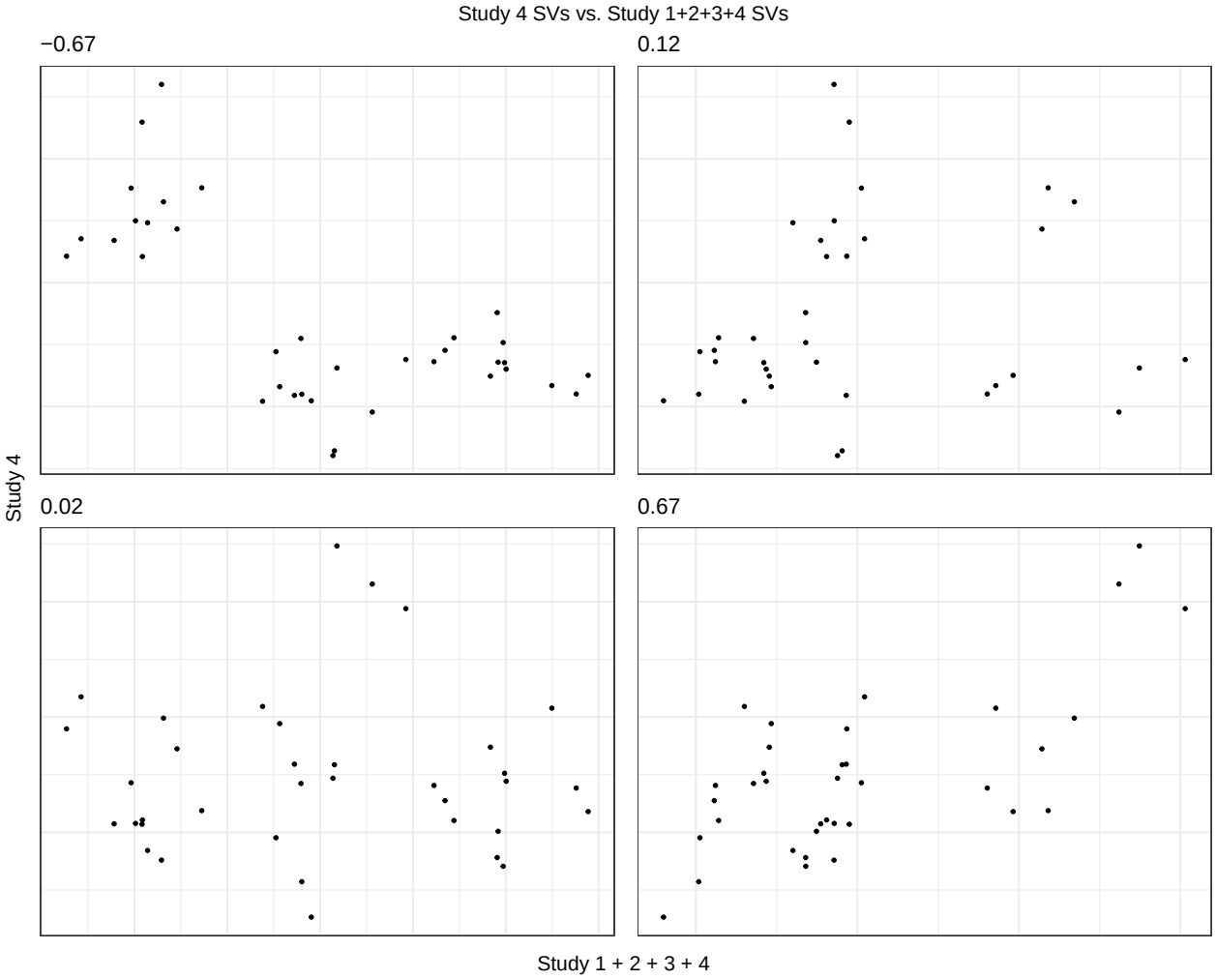
```
## [[1]]
## NULL
```



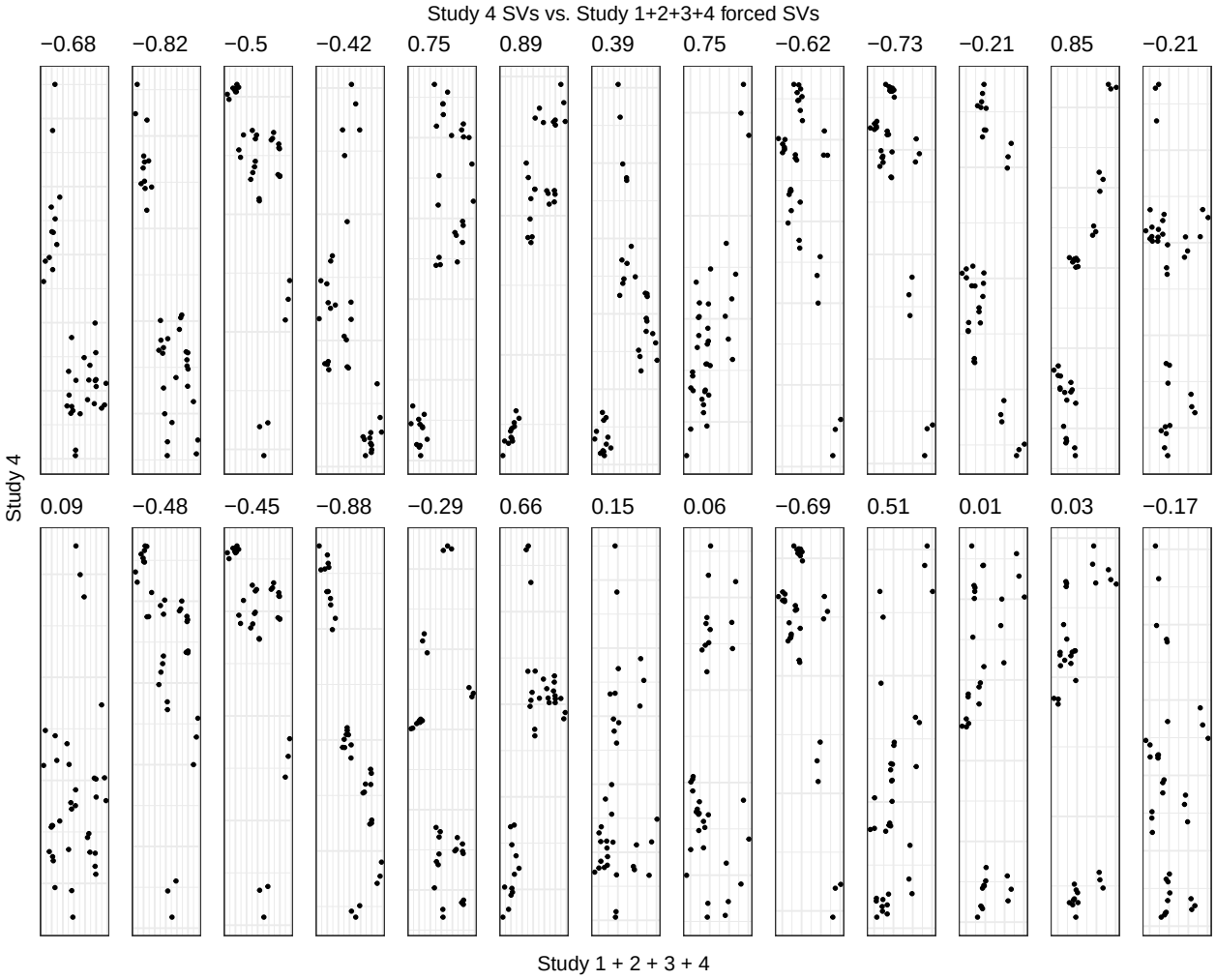

```
## [[1]]
## NULL
```


Study 3 SVs vs. Study 1+2+3+4 forced SVs
−0.04
−0.43
0.42
−0.01
−0.17
−0.33
−0.13
−0.05
0.28
−0.16
0.07
−0.02
0.16
Study 3
Study 1 + 2 + 3 + 4

```
## [[1]]
## NULL
```


Study 4 SVs vs. Study 1+2+3+4 SVs
−0.67
0.12
0.02
0.67
Study 4
Study 1 + 2 + 3 + 4

```
## [[1]]
## NULL
```


Study 4 SVs vs. Study 1+2+3+4 forced SVs
−0.68
−0.82
−0.5
−0.42
0.75
0.89
0.39
0.75
−0.62
−0.73
−0.21
0.85
−0.21
0.09
−0.48
−0.45
−0.88
−0.29
0.66
0.15
0.06
−0.69
0.51
0.01
0.03
−0.17
Study 4
Study 1 + 2 + 3 + 4

```
## [[1]]
## NULL
```