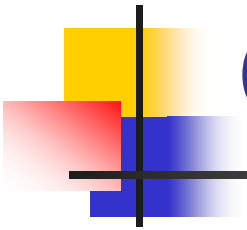
Visualization Plots for Applying CCANCOR to Bee cDNA Data

D. Noe, J. Zhou,
S. Rodriguez-Zas, X. He, B. Bailey



Outline

- Introduction
- Sample Analysis
  - Gene subset selection
  - Monte Carlo results

Introduction

Introduction

- Our data
  - cDNA microarray data from the honeybee experiment
    - Original data has 6 treatment levels h1, h2, ... , h6 and three bees
    - Loop design resulting in 20 arrays (i.e., 40 observations)
  - Data are normalized with dye, array, and trend effects removed
  - Our analysis considers only data from h2 and h3 (6 observations each)

Introduction

- Initial objective
  - Identify linear combinations of small groups of genes that are potentially useful in classifying bees into their age groups
- Method
  - Constrained canonical correlation (CCANCOR)

Introduction

- Potential drawback
  - CCANCOR is an iterative algorithm, requiring an initial set of coefficients to enable estimation
  - Solutions vary depending on the initial values
    - Coefficients of selected genes vary
    - Selected genes themselves may vary
- Goal
  - Examine the impact of initial values on CCANCOR results
  - Potentially provide scientist with multiple viable solutions for consideration

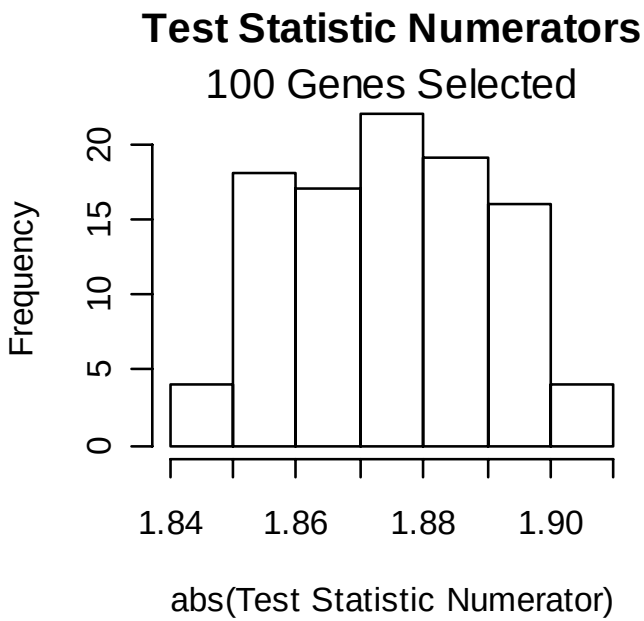
Sample Analysis

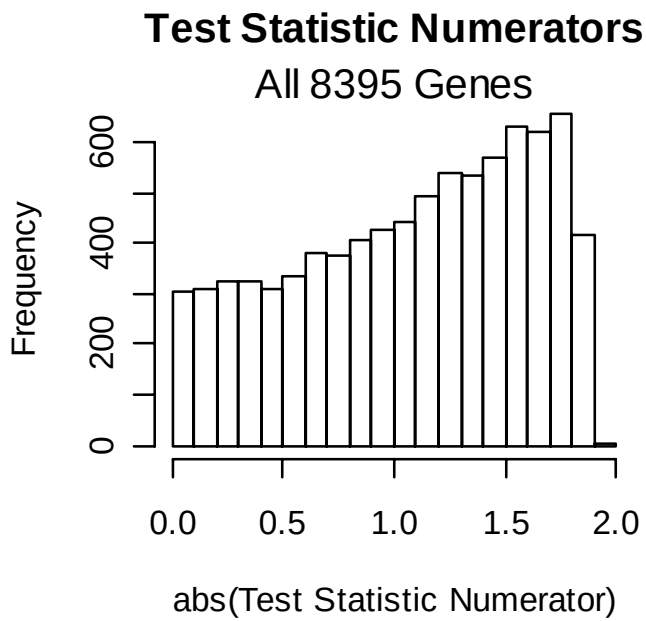
Sample Analysis

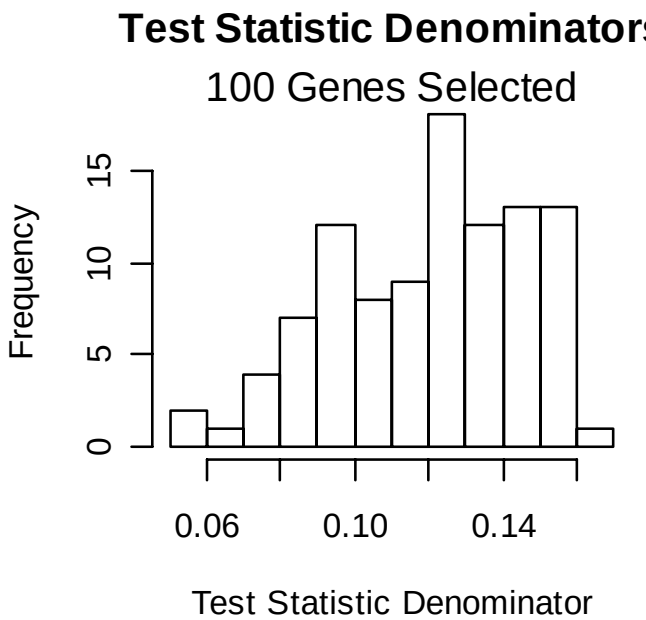
1. Reduce 8000+ genes to a more managable set of 100
   - Subset selected based on T-stat p-values
2. Use Monte Carlo methods to examine initial value effect
   - 1000 iterations
   - At each iteration:
     - Randomly assign equal, positive weights to 7 of the 100 genes; zero weights to all others
     - Record “Top 5” genes selected by CCANCOR procedure
   - Compare gene sets from each iteration to provide scientist with qualitative results

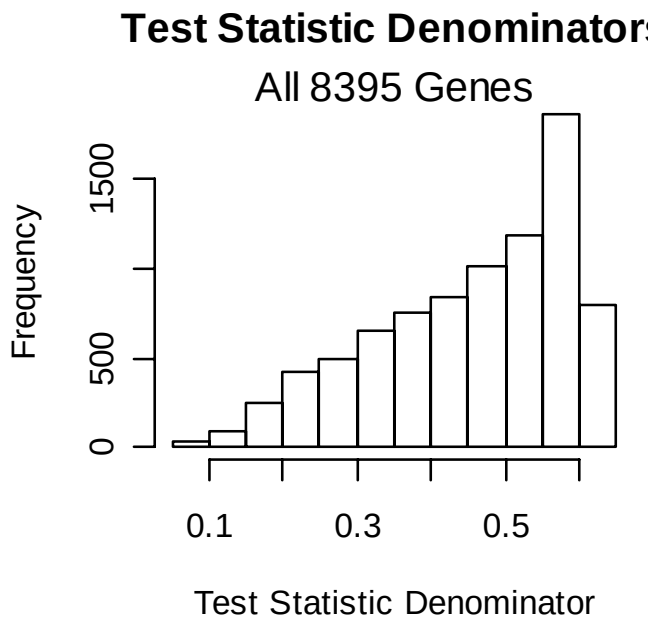
Step 1 – Narrow to 100 genes

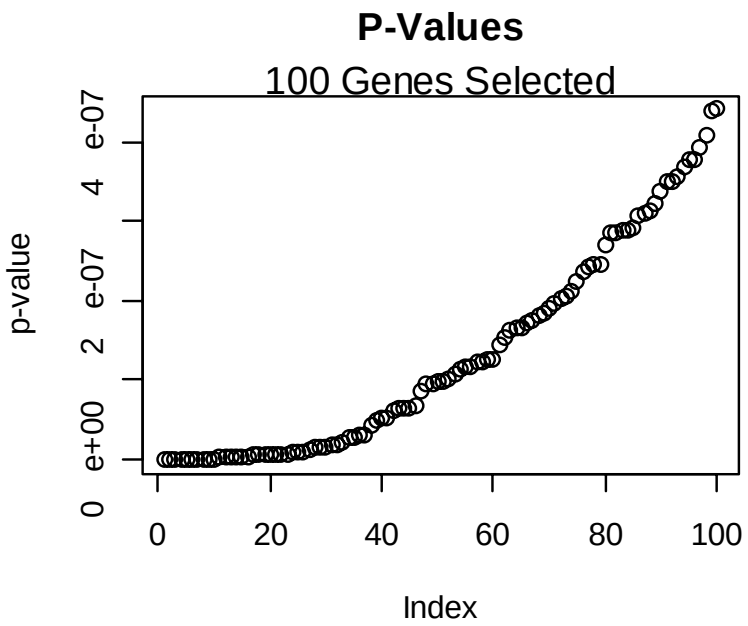
- Procedure
  - Perform 2-sample t-test for each gene
  - Renumber genes in increasing order of p-value
  - Select genes with the 100 smallest p-values for investigation
- Related plots
  - T-statistic component and p-value plots for selected 100 genes
  - T-statistic component comparison between 100 selected genes and all 8395 genes

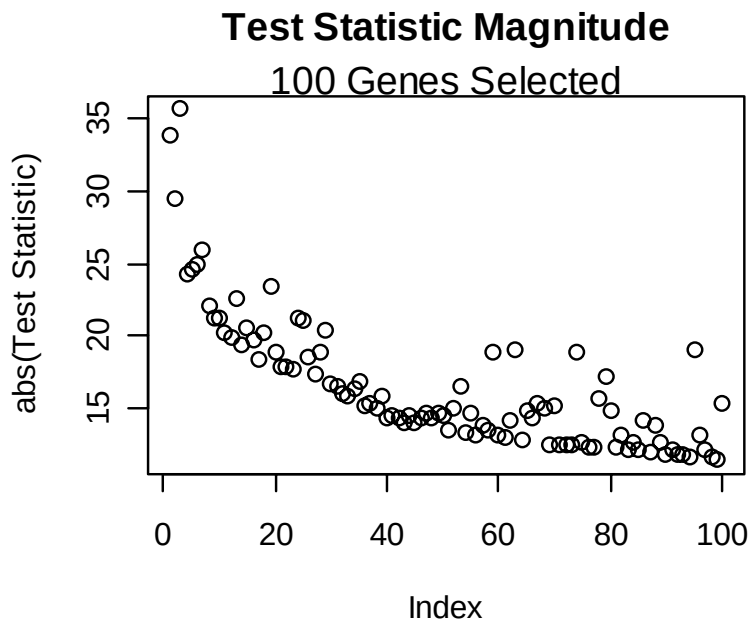

Test Statistic Numerators
100 Genes Selected
Frequency
0
5
10
15
20
1.84
1.86
1.88
1.90
abs(Test Statistic Numerator)

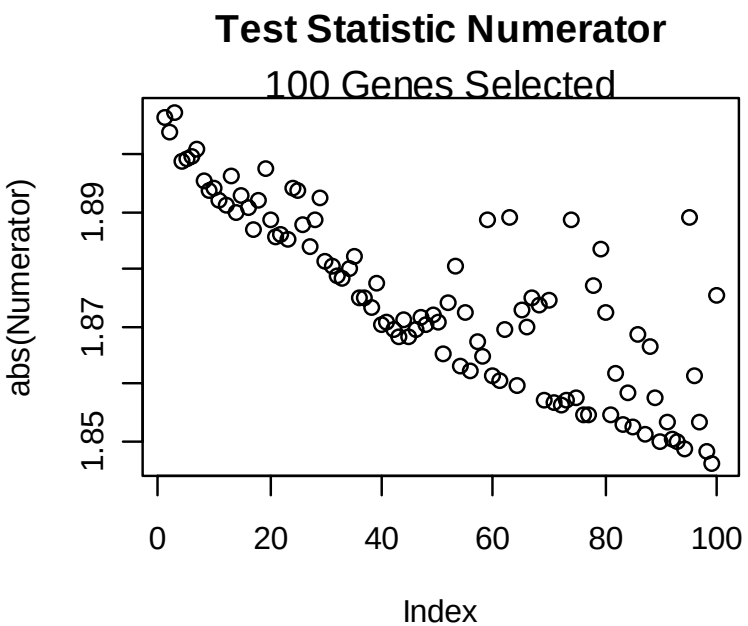

Test Statistic Numerators
All 8395 Genes
Frequency
0
200
400
600
0.0
0.5
1.0
1.5
2.0
abs(Test Statistic Numerator)

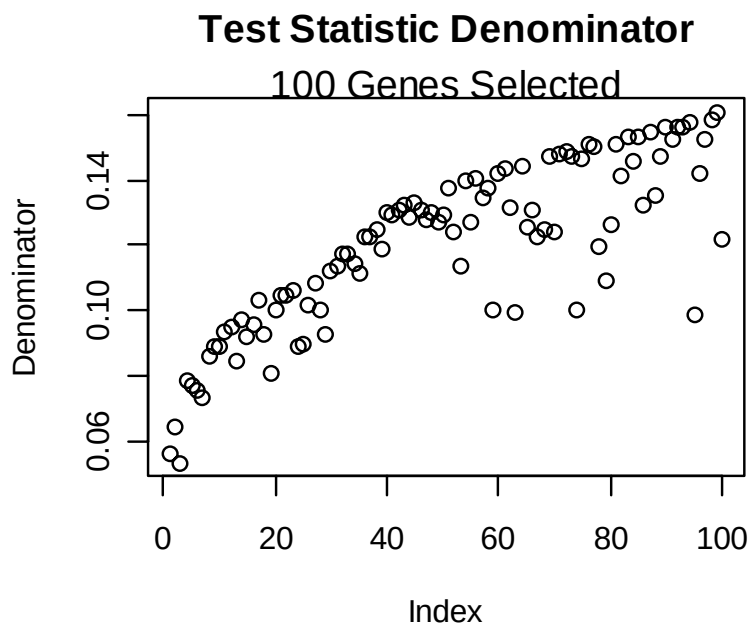

Test Statistic Denominators
100 Genes Selected
Frequency
0
5
10
15
0.06
0.10
0.14
Test Statistic Denominator


Test Statistic Denominators
All 8395 Genes
Frequency
0
500
1500
0.1
0.3
0.5
Test Statistic Denominator


P-Values
100 Genes Selected
p-value
0 e+00
2 e-07
4 e-07
0
20
40
60
80
100
Index


Test Statistic Magnitude
100 Genes Selected
abs(Test Statistic)
15
20
25
30
35
0
20
40
60
80
100
Index


Test Statistic Numerator
100 Genes Selected
abs(Numerator)
1.85
1.87
1.89
0
20
40
60
80
100
Index

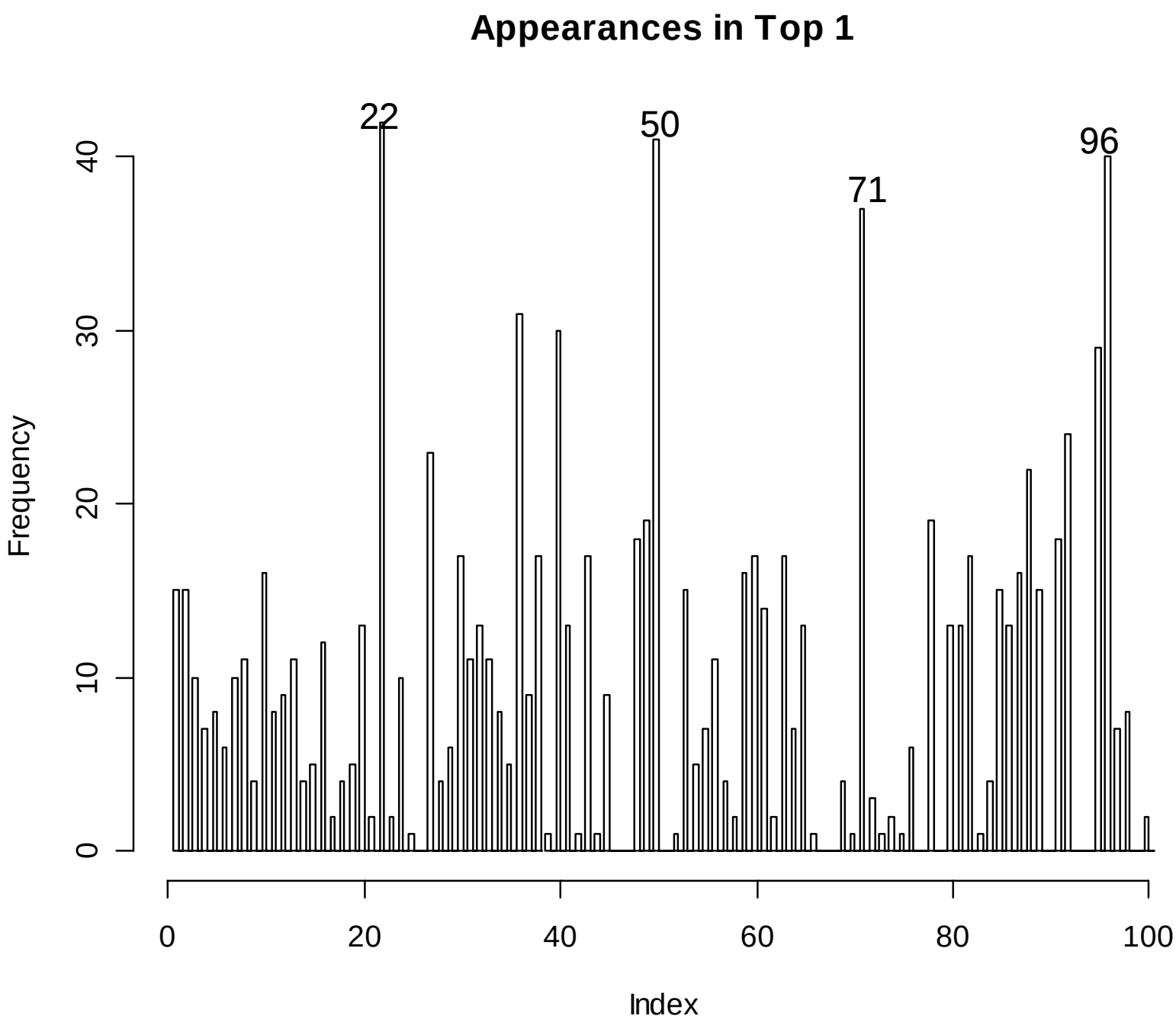

Test Statistic Denominator
100 Genes Selected
Denominator
0.06
0.10
0.14
0
20
40
60
80
100
Index

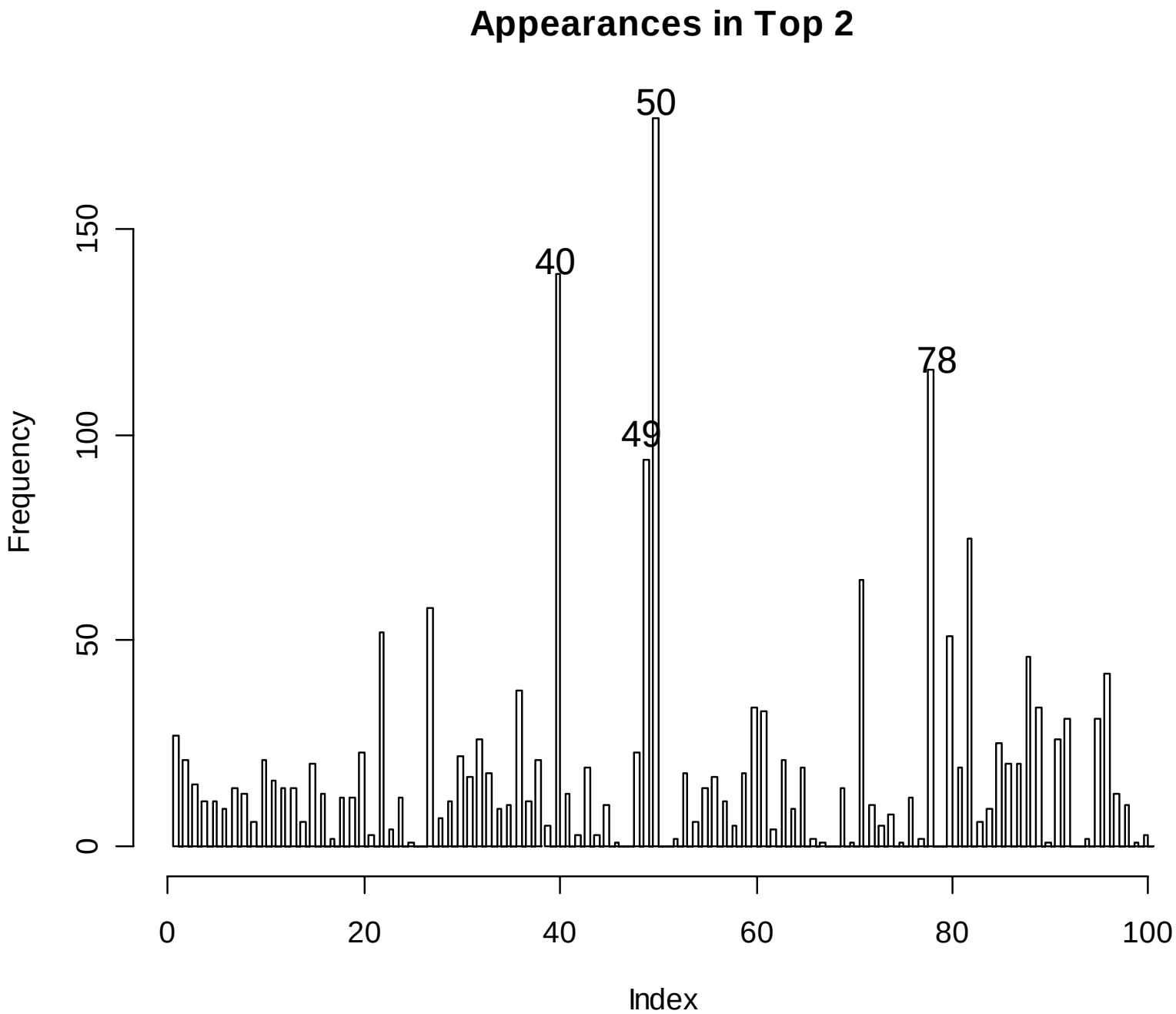
Step 2 – Monte Carlo

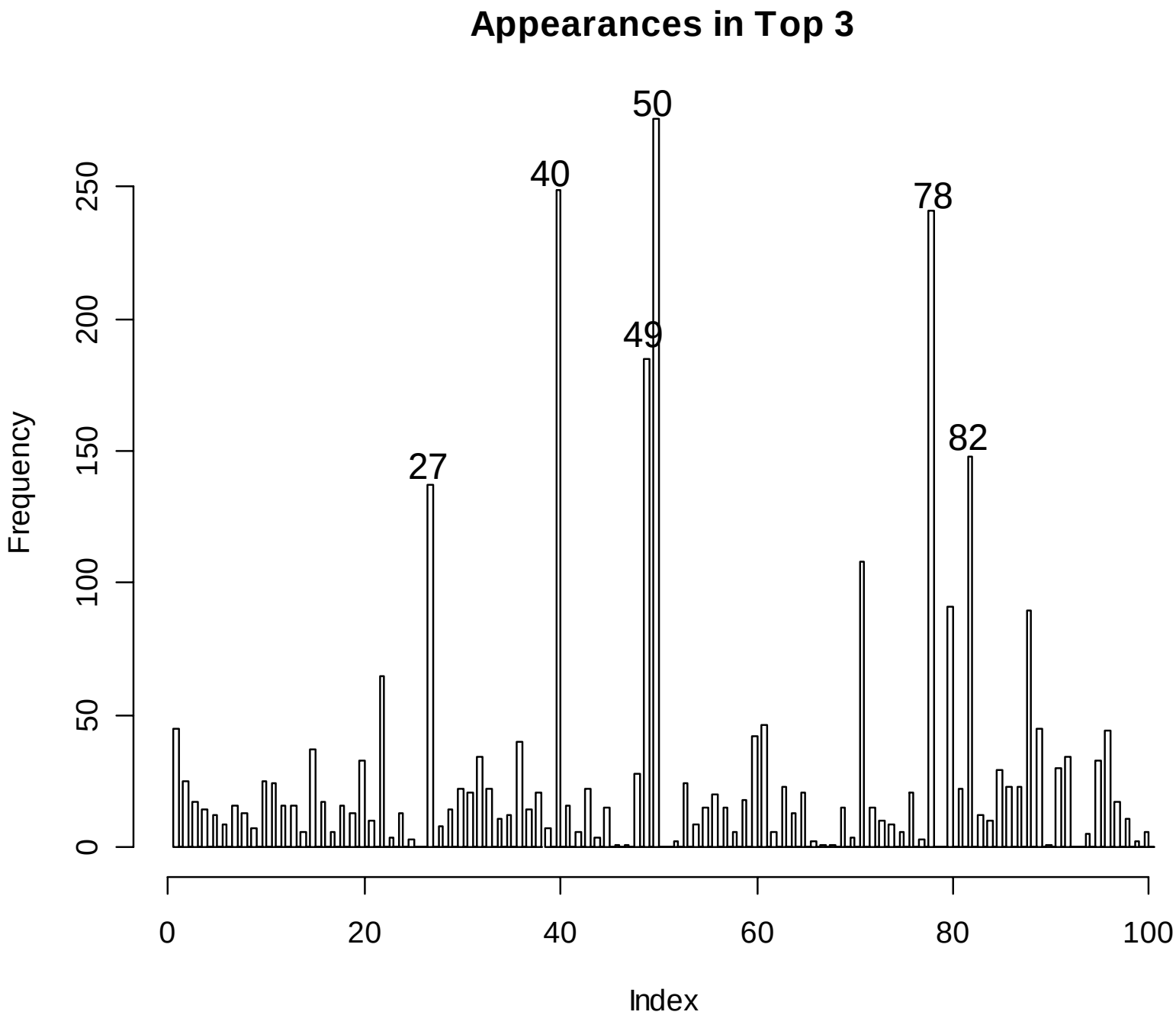
- For each of 1000 iterations,
  - Randomly select 7 genes to receive positive initial weight
    - Based on observation, 7 appears to be the inherent rank for this data
  - Record the 5 genes CCANCOR identifies as most important (i.e., having the largest coefficients in magnitude)

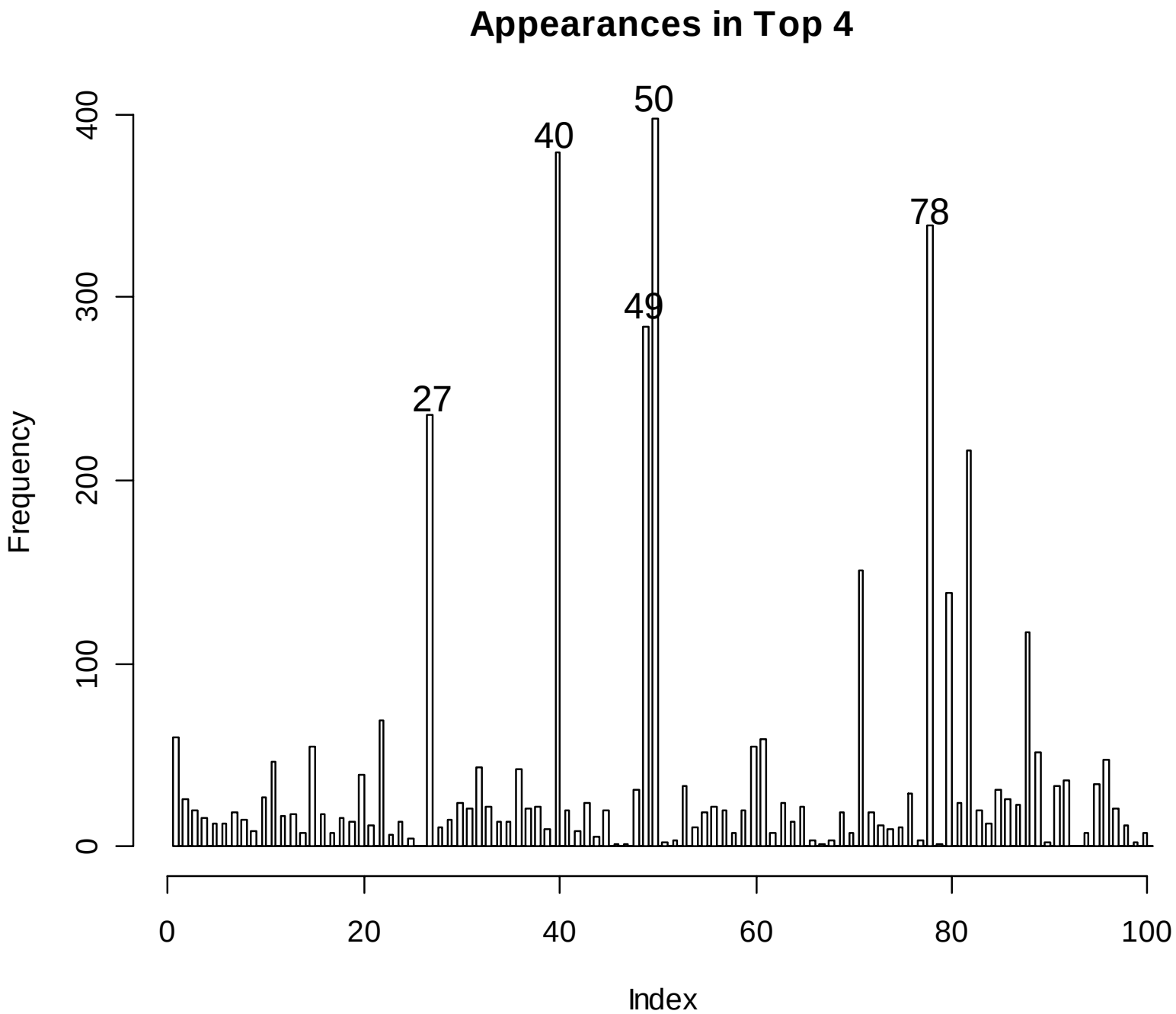
Step 2 – Monte Carlo

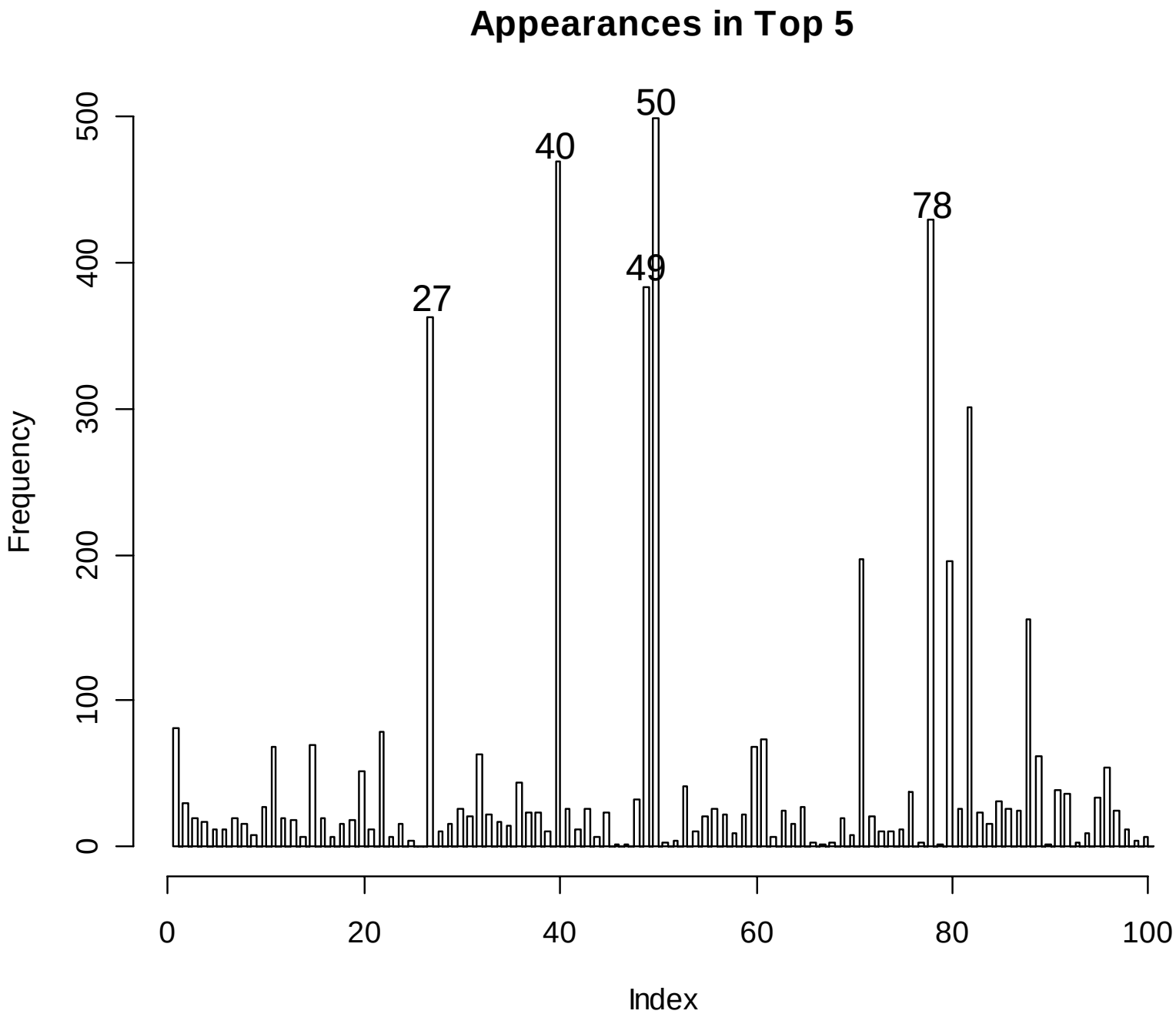
- Related plots
  - Assessing importance of individual genes
    - Winners plot – How often was each gene assigned the largest coefficient?
    - Top n plot – How often was each gene among the top n in magnitude


Appearances in Top 1
22
50
71
96
Frequency
0
10
20
30
40
0
20
40
60
80
100
Index


Appearances in Top 2
50
40
78
49
Frequency
0
50
100
150
0
20
40
60
80
100
Index

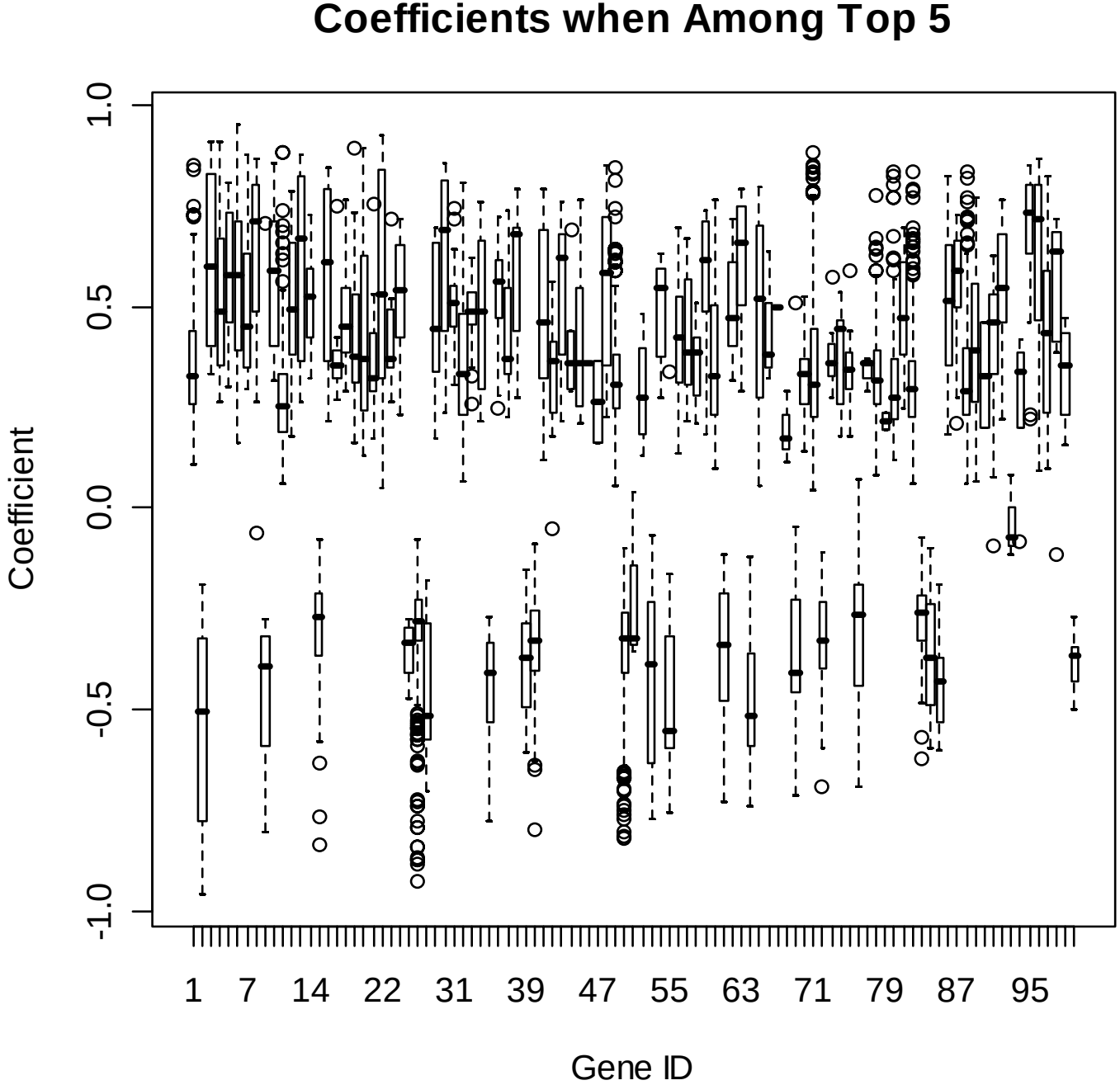

Appearances in Top 3
50
40
78
49
27
82
Frequency
0
50
100
150
200
250
0
20
40
60
80
100
Index

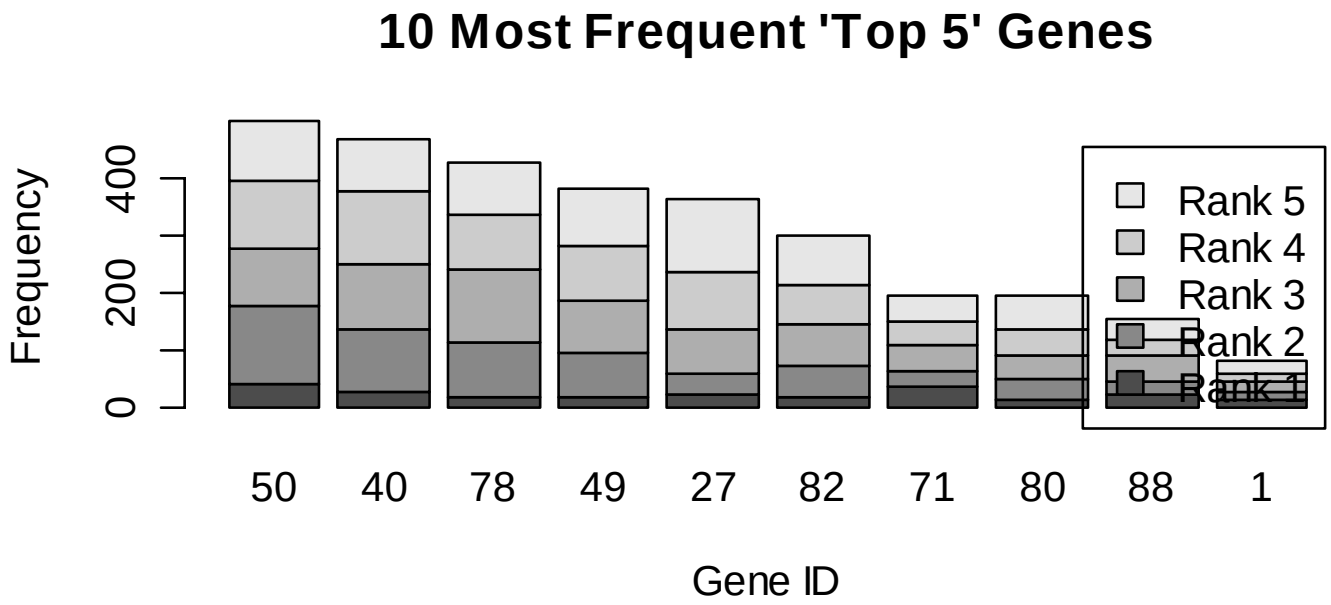

Appearances in Top 4
Frequency
0
100
200
300
400
27
40
49
50
78
0
20
40
60
80
100
Index

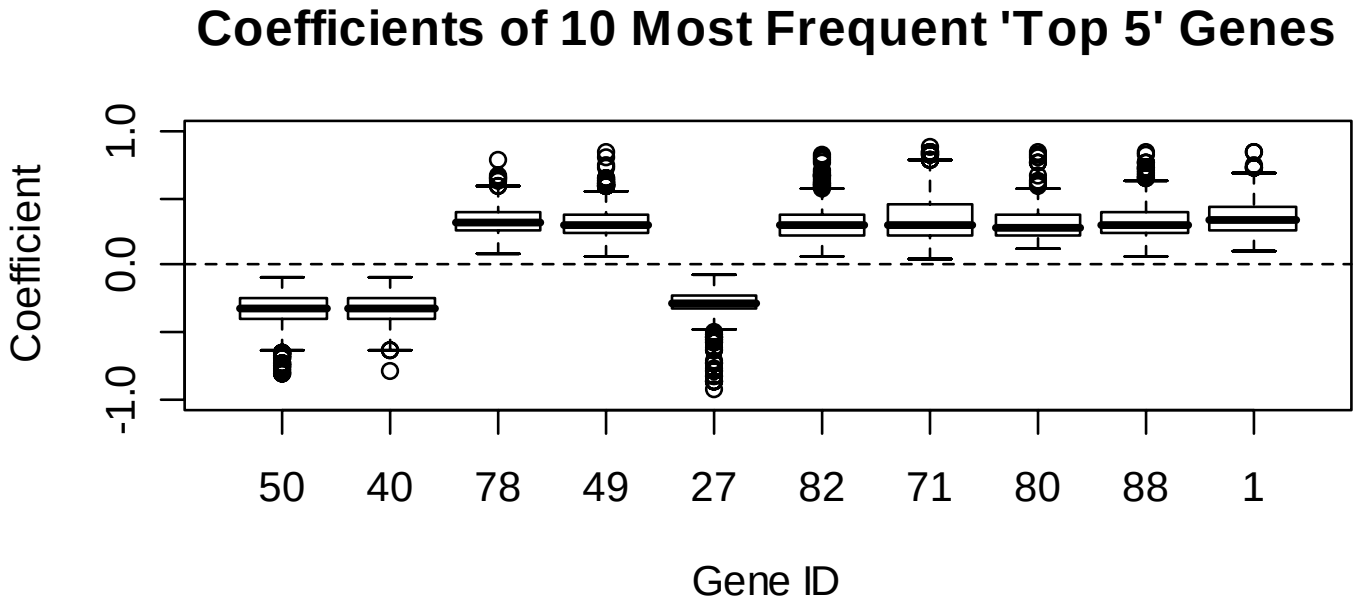

Appearances in Top 5
Frequency
0
100
200
300
400
500
27
40
49
50
78
0
20
40
60
80
100
Index

Step 2 – Monte Carlo

- Related plots (continued)
  - Assessing importance of individual genes
    - Coefficients plot – What weights are assigned to each gene when it appears in the top 5?
    - Top genes plot
      - Which genes appear in the top 5 most often?
      - What are their importance rank when they appear?
      - What are their coefficients when they appear?

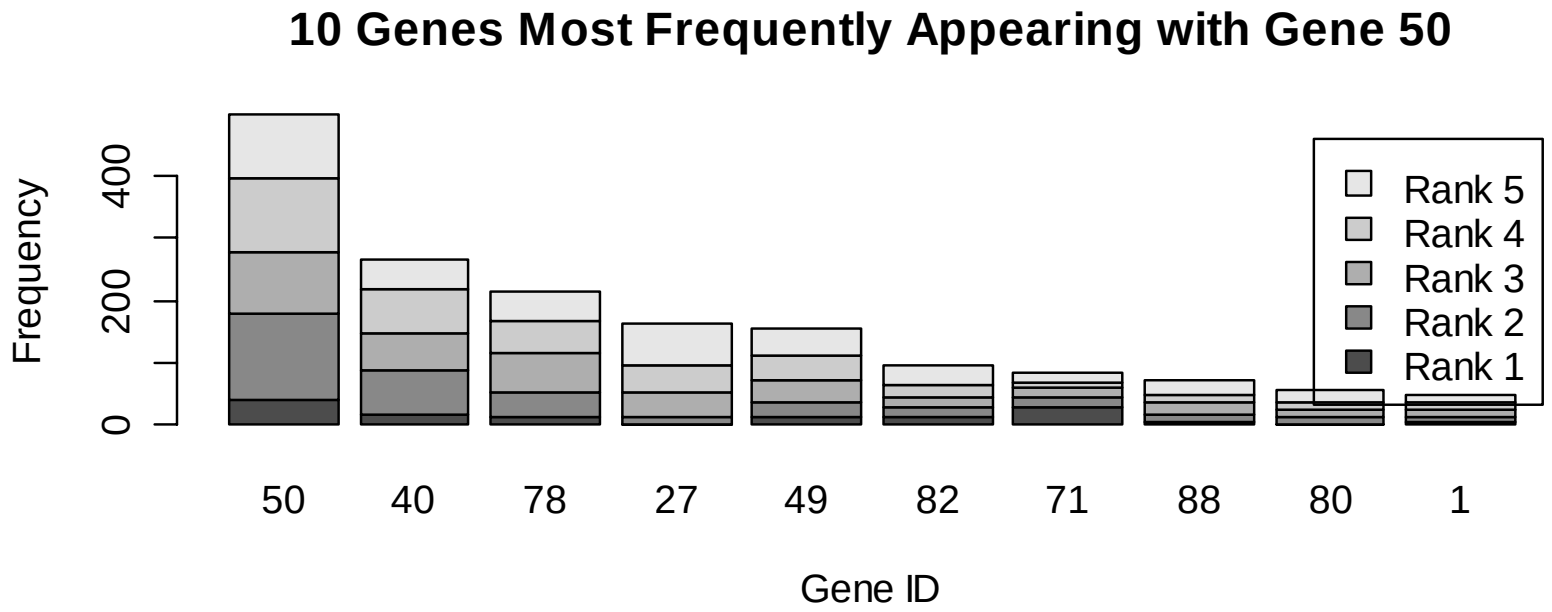

Coefficients when Among Top 5
Coefficient
1.0
0.5
0.0
-0.5
-1.0
1
7
14
22
31
39
47
55
63
71
79
87
95
Gene ID

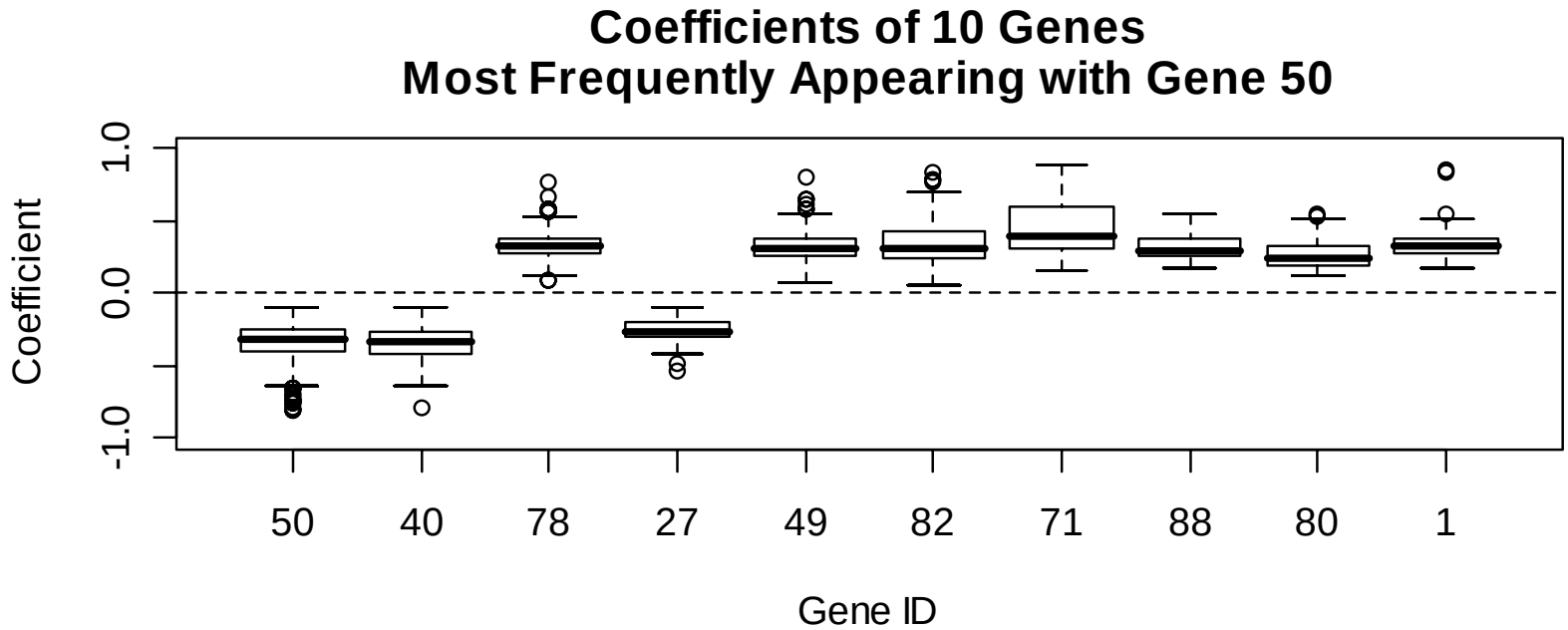

10 Most Frequent 'Top 5' Genes
Frequency
0
200
400
Rank 5
Rank 4
Rank 3
Rank 2
Rank 1
50
40
78
49
27
82
71
80
88
1
Gene ID

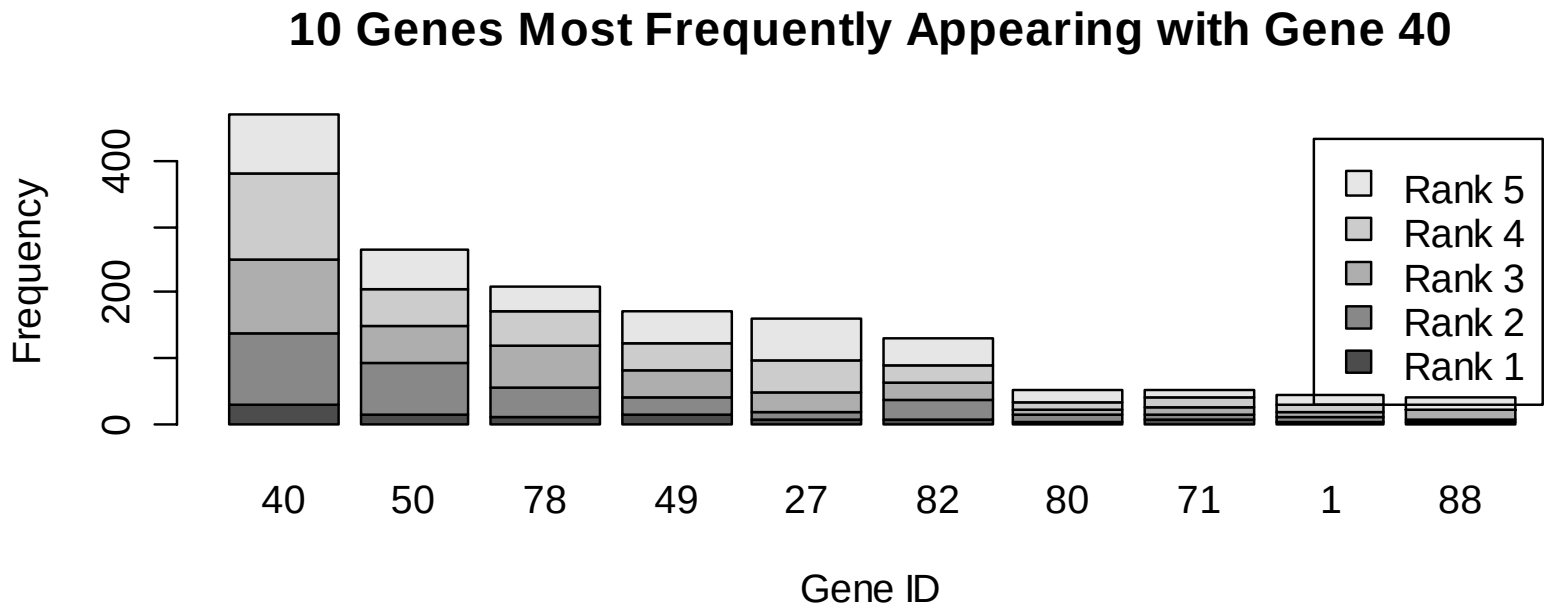

Coefficients of 10 Most Frequent 'Top 5' Genes
Coefficient
-1.0
0.0
1.0
50
40
78
49
27
82
71
80
88
1
Gene ID

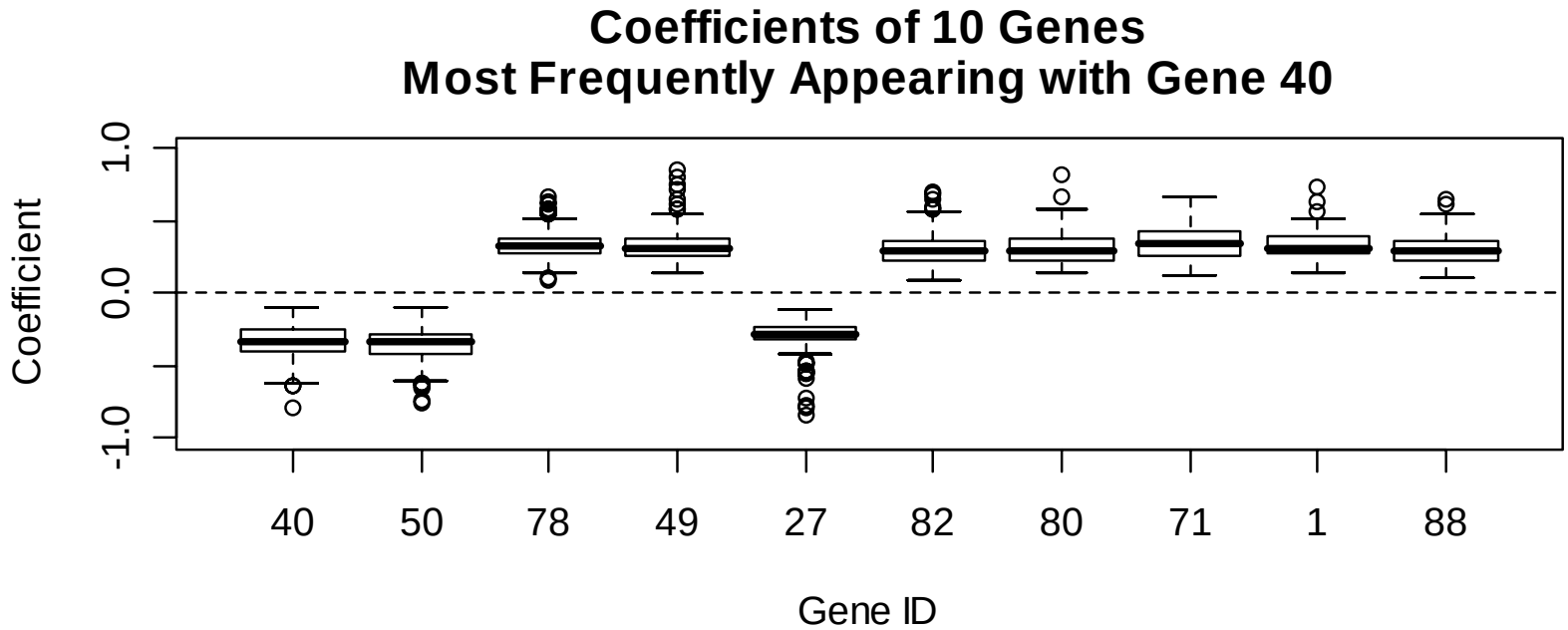
Step 2 – Monte Carlo

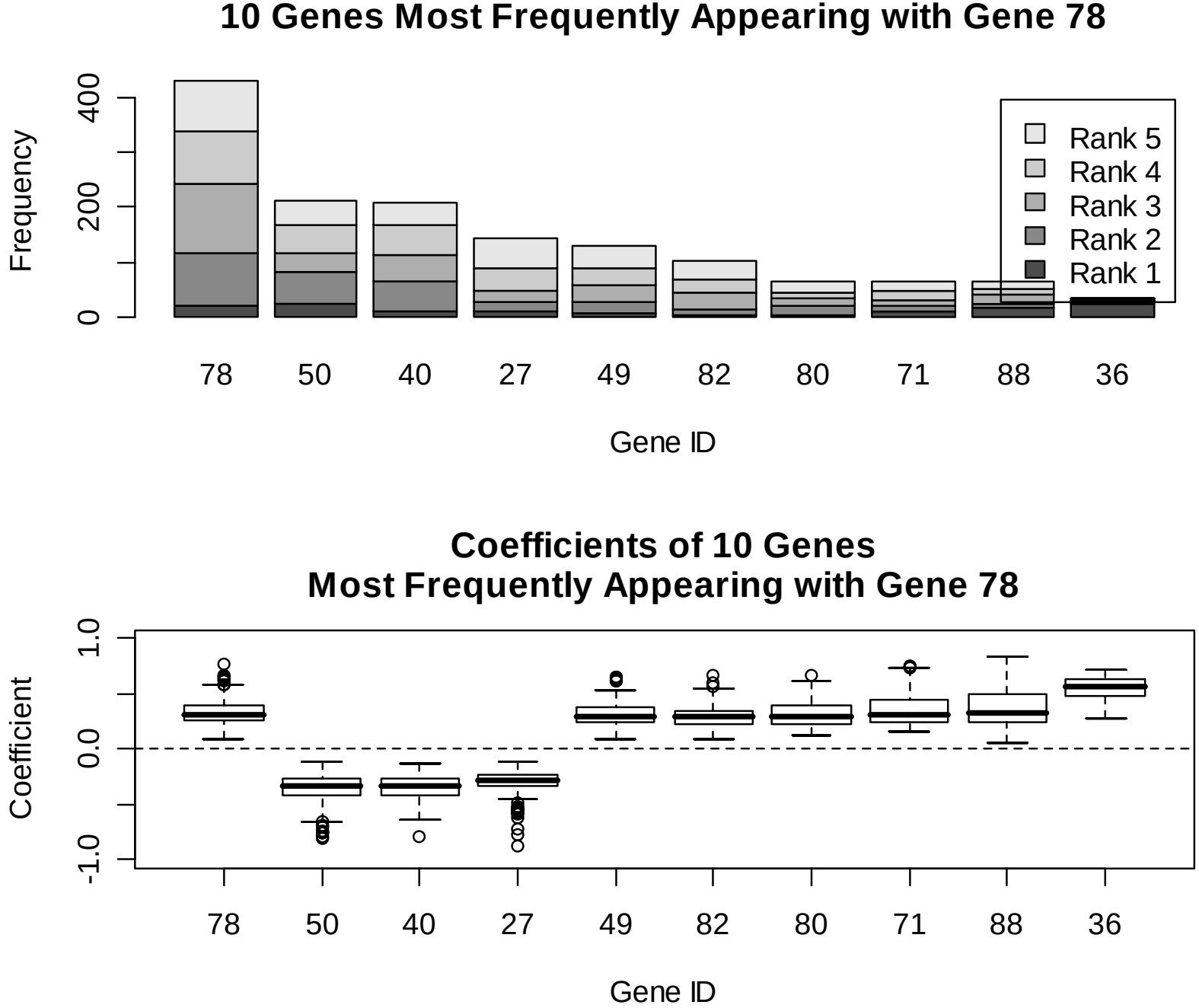
- Related plots (continued)
  - Assessing relationships between genes
    - Top accomplices plot
      - Which genes appear in the top 5 with a particular gene most often?
      - What are their importance rank when they appear?
      - What are their coefficients when they appear?

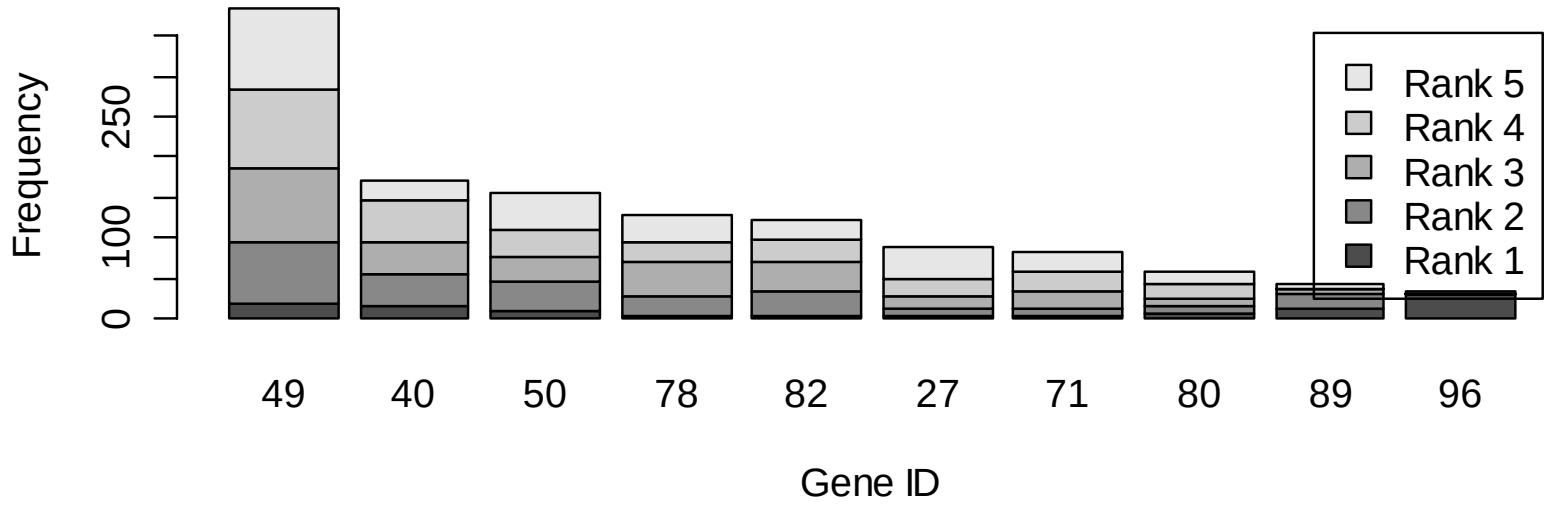

10 Genes Most Frequently Appearing with Gene 50
Frequency
0
200
400
Rank 5
Rank 4
Rank 3
Rank 2
Rank 1
50
40
78
27
49
82
71
88
80
1
Gene ID

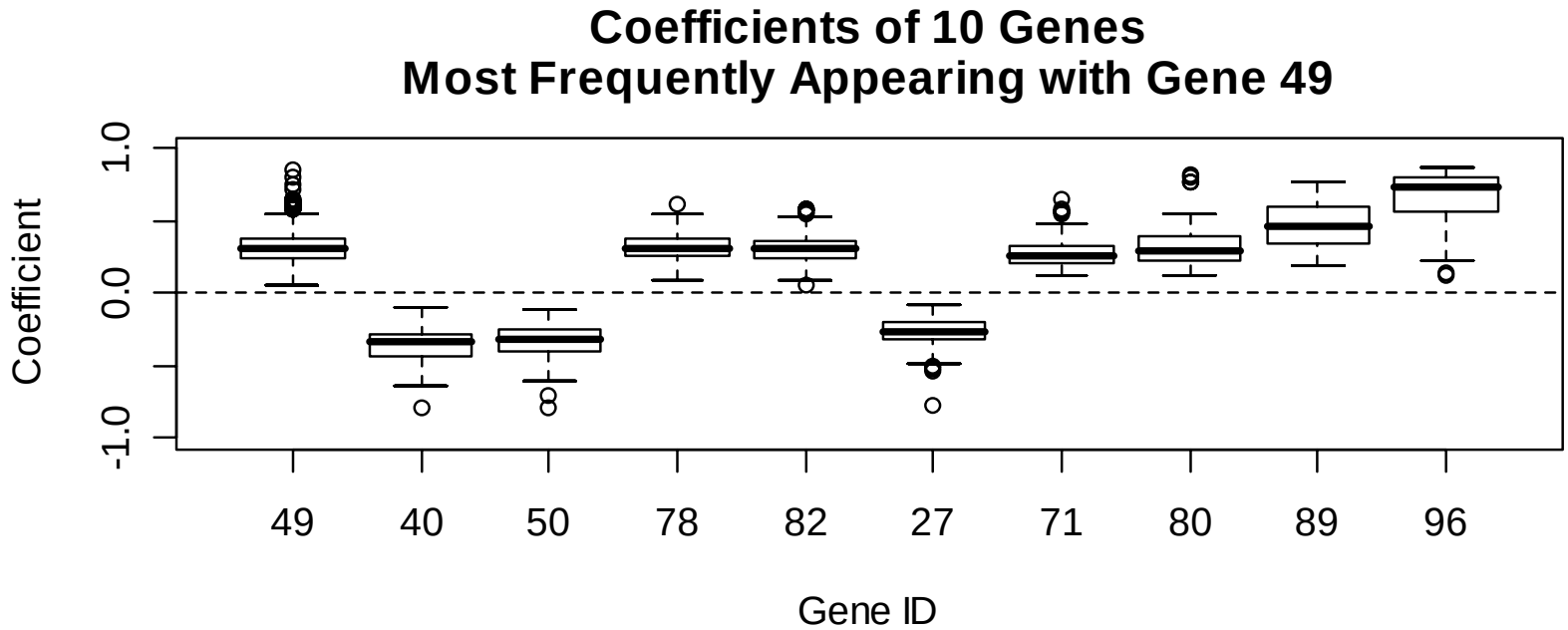

Coefficients of 10 Genes
Most Frequently Appearing with Gene 50
Coefficient
-1.0
0.0
1.0
50
40
78
27
49
82
71
88
80
1
Gene ID

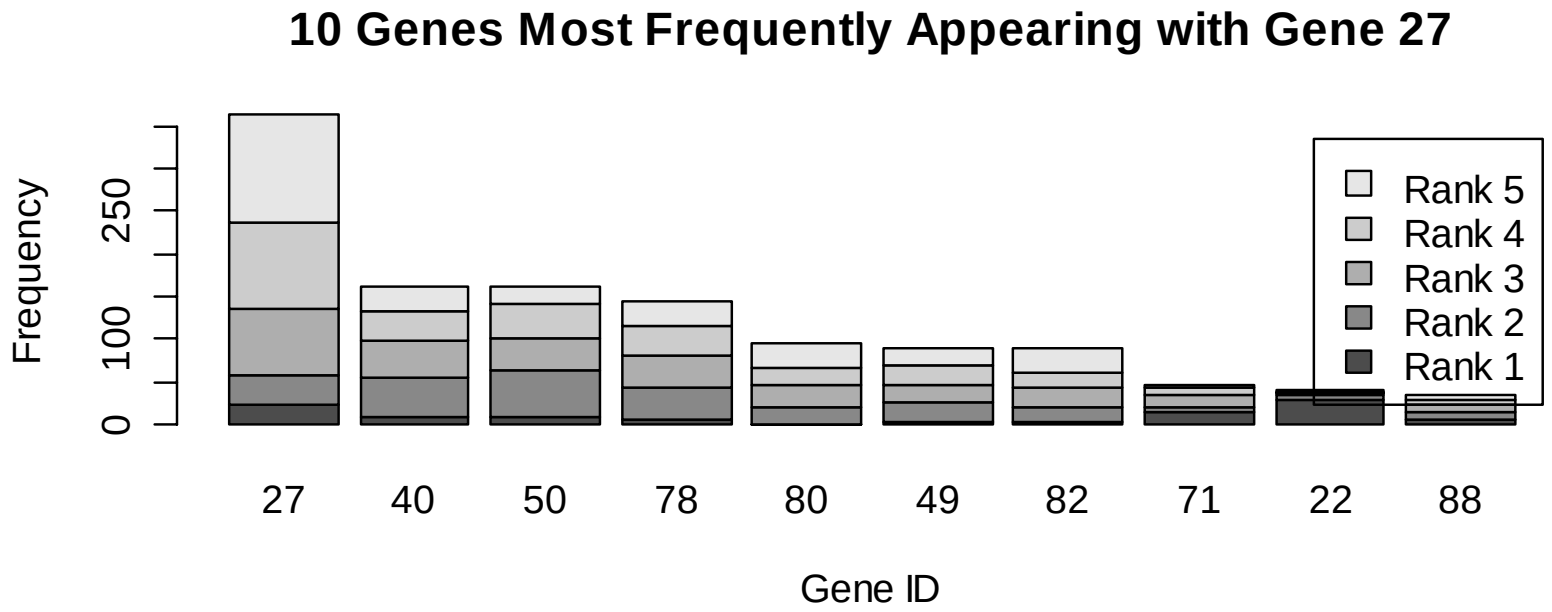

10 Genes Most Frequently Appearing with Gene 40
Frequency
0
200
400
Rank 5
Rank 4
Rank 3
Rank 2
Rank 1
40
50
78
49
27
82
80
71
1
88
Gene ID

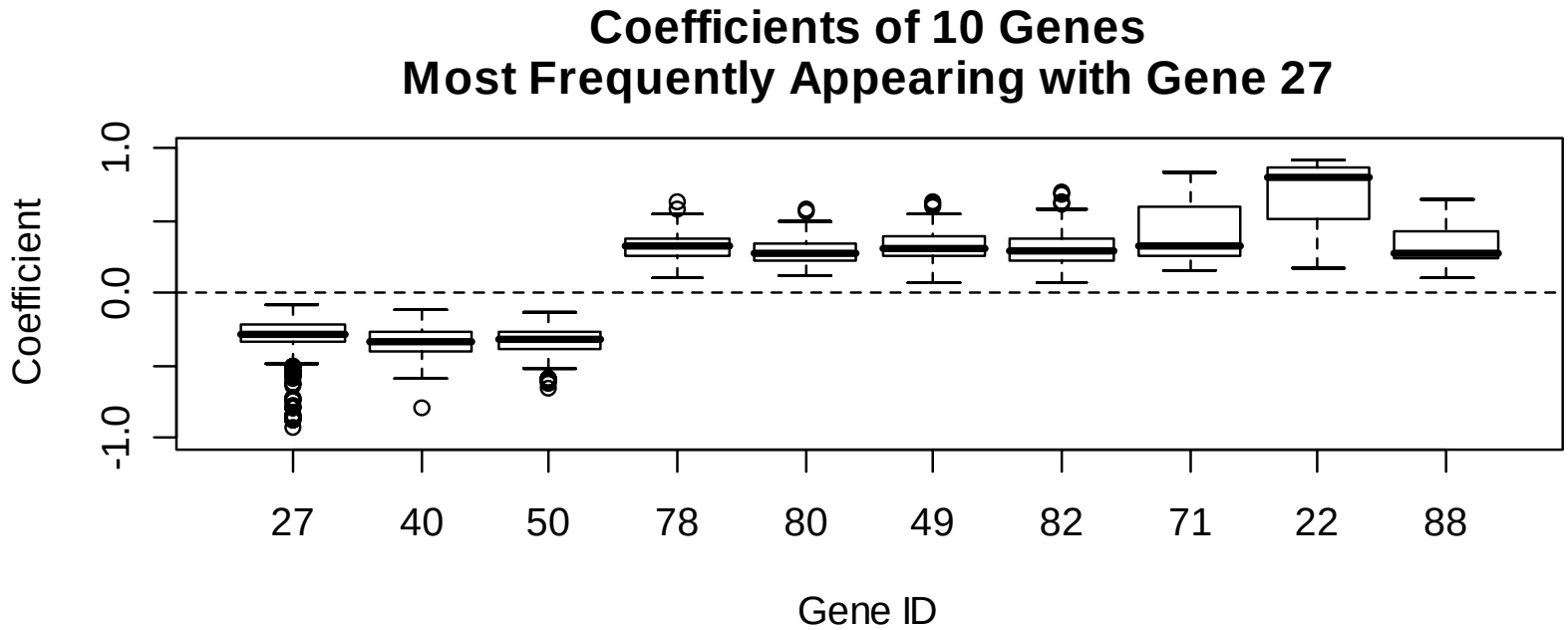

Coefficients of 10 Genes
Most Frequently Appearing with Gene 40
Coefficient
-1.0
0.0
1.0
40
50
78
49
27
82
80
71
1
88
Gene ID


10 Genes Most Frequently Appearing with Gene 78
Frequency
0
200
400
Rank 5
Rank 4
Rank 3
Rank 2
Rank 1
78
50
40
27
49
82
80
71
88
36
Gene ID
Coefficients of 10 Genes
Most Frequently Appearing with Gene 78
Coefficient
-1.0
0.0
1.0
78
50
40
27
49
82
80
71
88
36
Gene ID


10 Genes Most Frequently Appearing with Gene 49
Frequency
0
100
250
Rank 5
Rank 4
Rank 3
Rank 2
Rank 1
49
40
50
78
82
27
71
80
89
96
Gene ID


Coefficients of 10 Genes
Most Frequently Appearing with Gene 49
Coefficient
1.0
0.0
-1.0
49
40
50
78
82
27
71
80
89
96
Gene ID

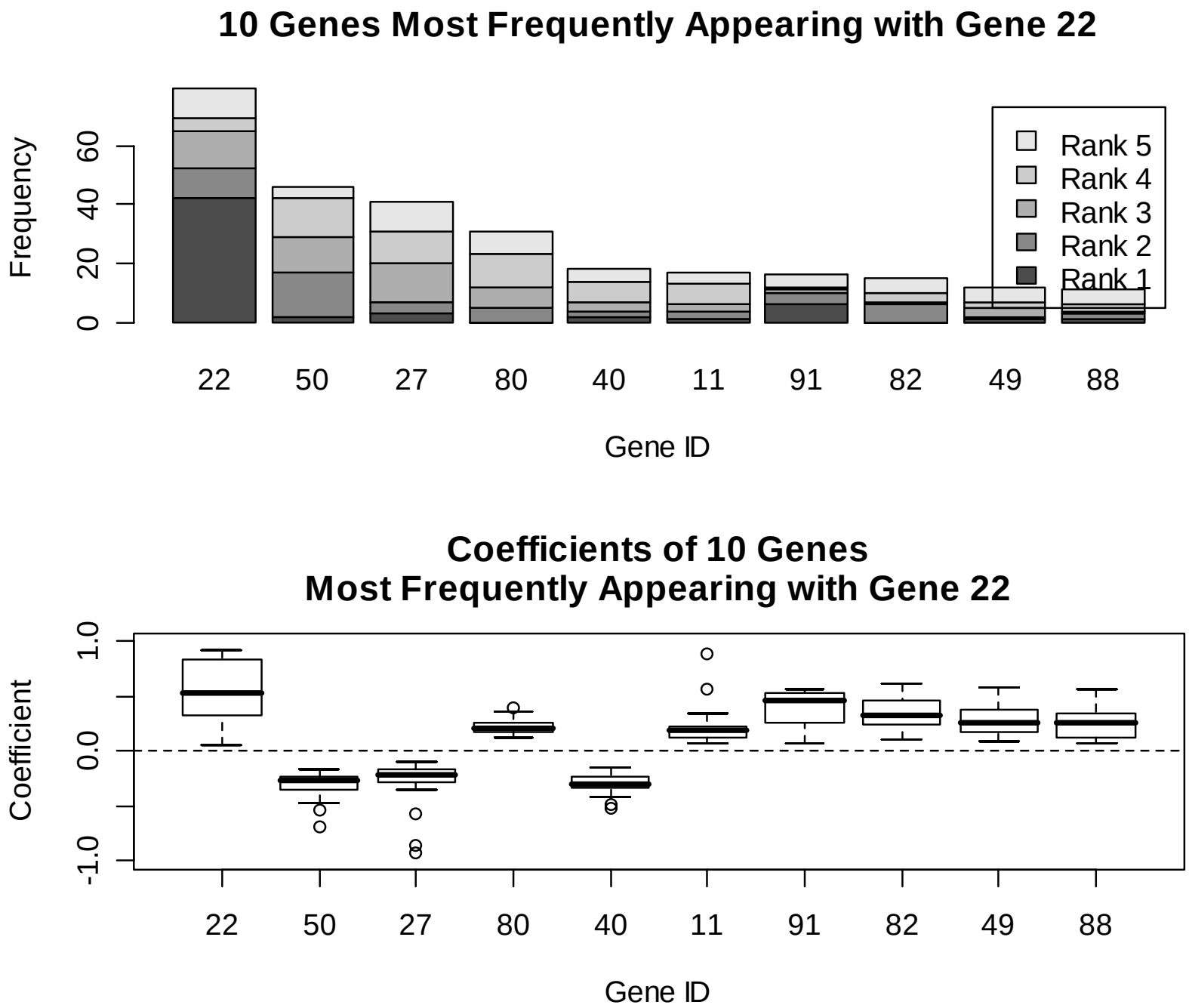

10 Genes Most Frequently Appearing with Gene 27
Frequency
0
100
250
Rank 5
Rank 4
Rank 3
Rank 2
Rank 1
27
40
50
78
80
49
82
71
22
88
Gene ID

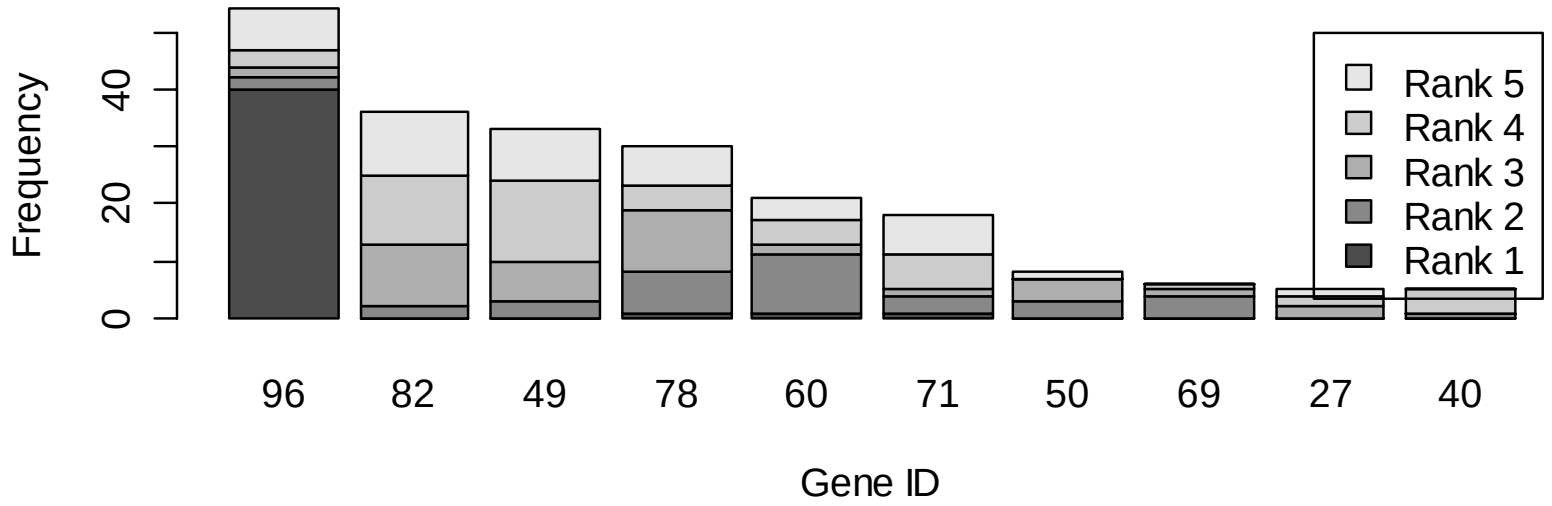

Coefficients of 10 Genes
Most Frequently Appearing with Gene 27
Coefficient
-1.0
0.0
1.0
27
40
50
78
80
49
82
71
22
88
Gene ID

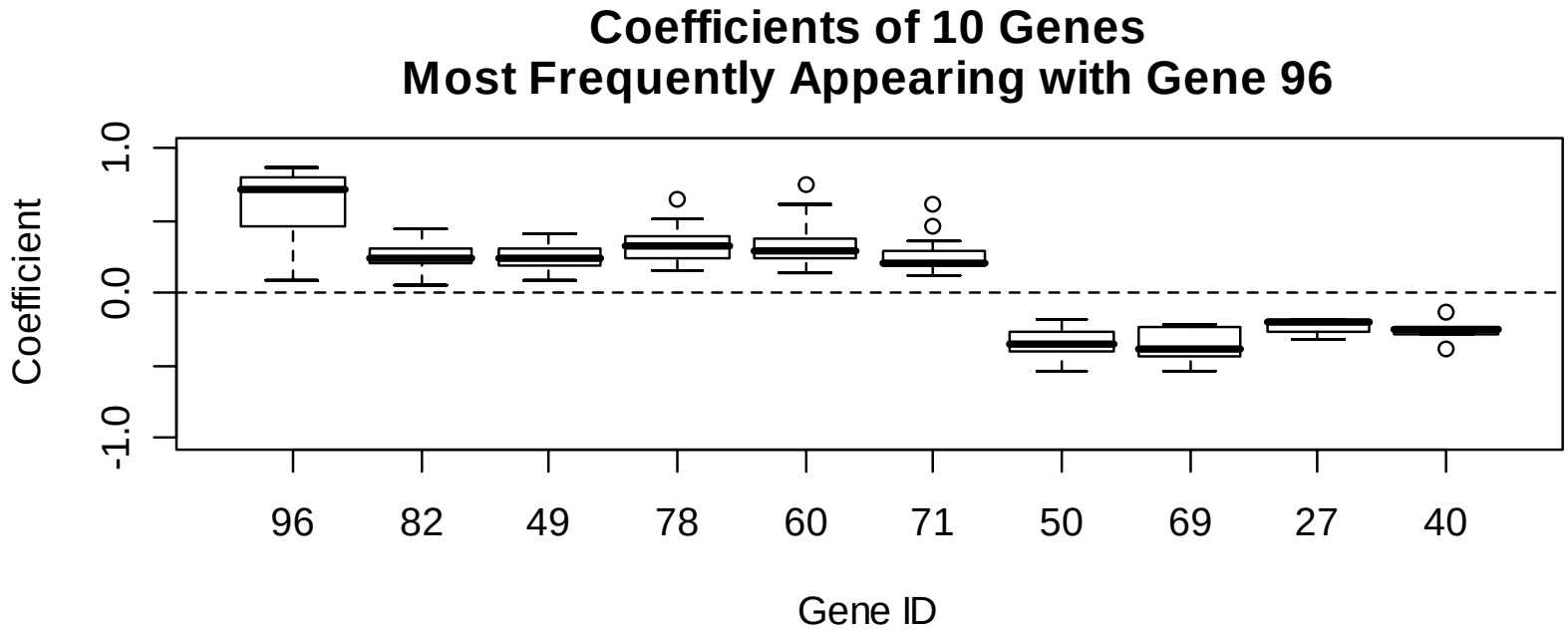
Step 2 – Monte Carlo

- Observations
  - Genes appearing in the top 5 frequently also appear *together* (pairwise) frequently
    - Genes and original ID, plus “known” functions
      - 50 → #7325
      - 40 → #8833
      - 78 → #5048
      - 49 → #5697
      - 27 → #333
        - Structural molecule activity, intracellular protein transport, synaptic transmission, neurotransmitter secretion, neuromuscular junction development, synaptic vesicle, synaptic vesicle priming, vesicle-mediated transport, transmission of nerve impulse
      - 82 → #2397

- Now will examine accomplices with other frequent “winners” that are less frequent among the top 5


10 Genes Most Frequently Appearing with Gene 22
Frequency
0
20
40
60
Rank 5
Rank 4
Rank 3
Rank 2
Rank 1
22
50
27
80
40
11
91
82
49
88
Gene ID
Coefficients of 10 Genes
Most Frequently Appearing with Gene 22
Coefficient
1.0
0.0
-1.0
22
50
27
80
40
11
91
82
49
88
Gene ID

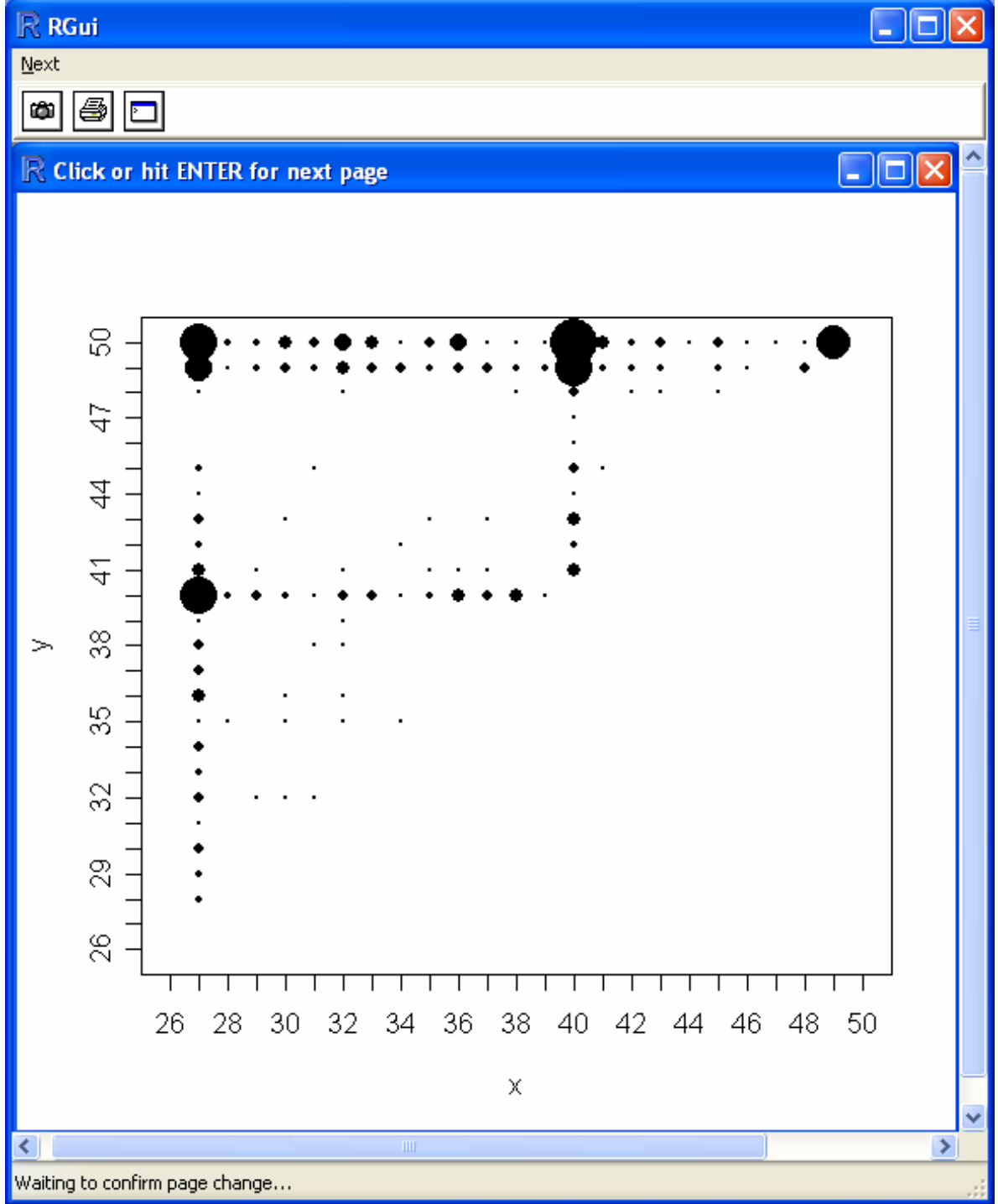

10 Genes Most Frequently Appearing with Gene 96
Frequency
0
20
40
Rank 5
Rank 4
Rank 3
Rank 2
Rank 1
96
82
49
78
60
71
50
69
27
40
Gene ID


Coefficients of 10 Genes
Most Frequently Appearing with Gene 96
Coefficient
1.0
0.0
-1.0
96
82
49
78
60
71
50
69
27
40
Gene ID

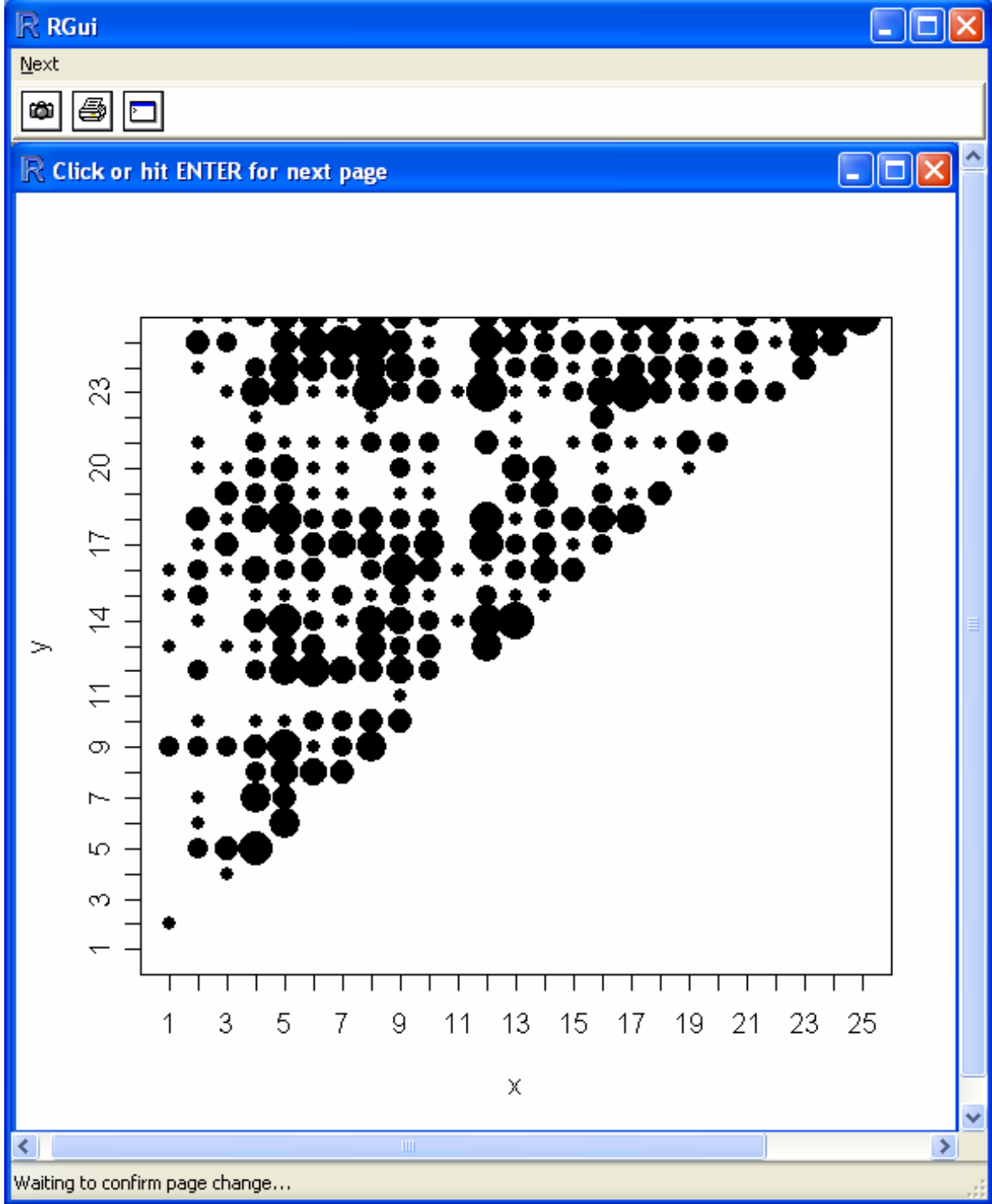
Step 2 – Monte Carlo

- Related plots (continued)
  - Assessing relationships between genes
    - Bubble Plots
      - Which genes appear together (pairwise) in the top 5 most often?
      - Which genes appear together (pairwise) in the bottom 5 most often?

Top 5 Bubbleplot



Bottom 5
Bubbleplot

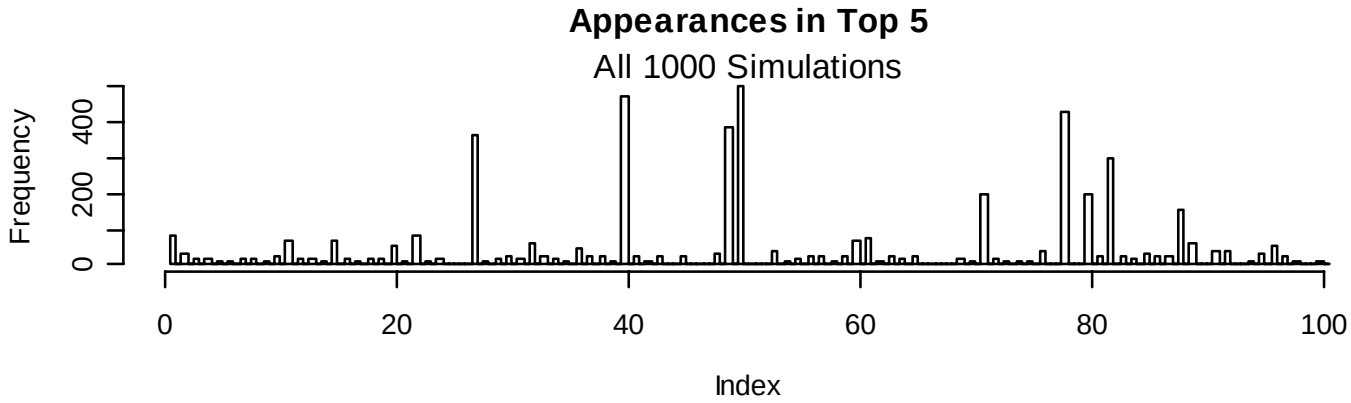


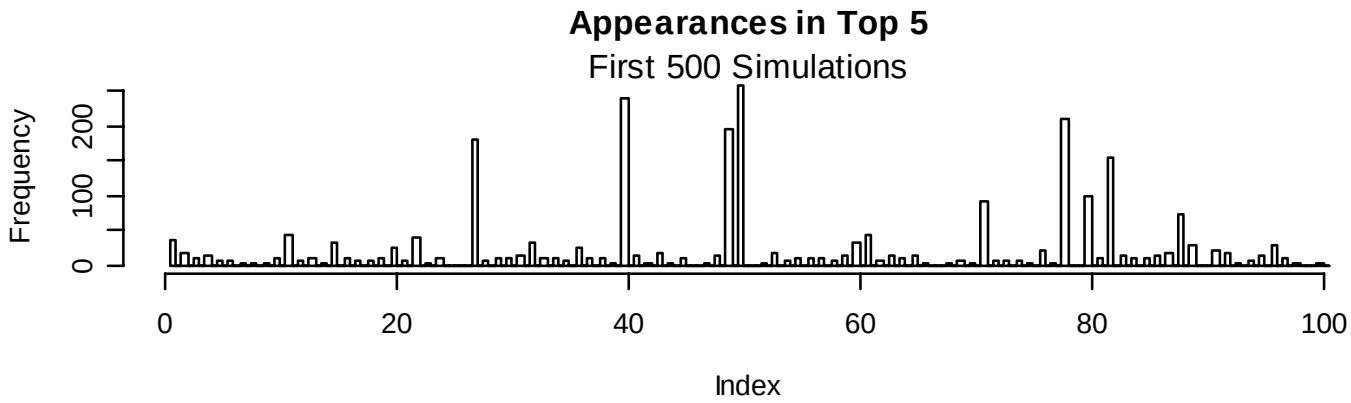
Step 2 – Monte Carlo

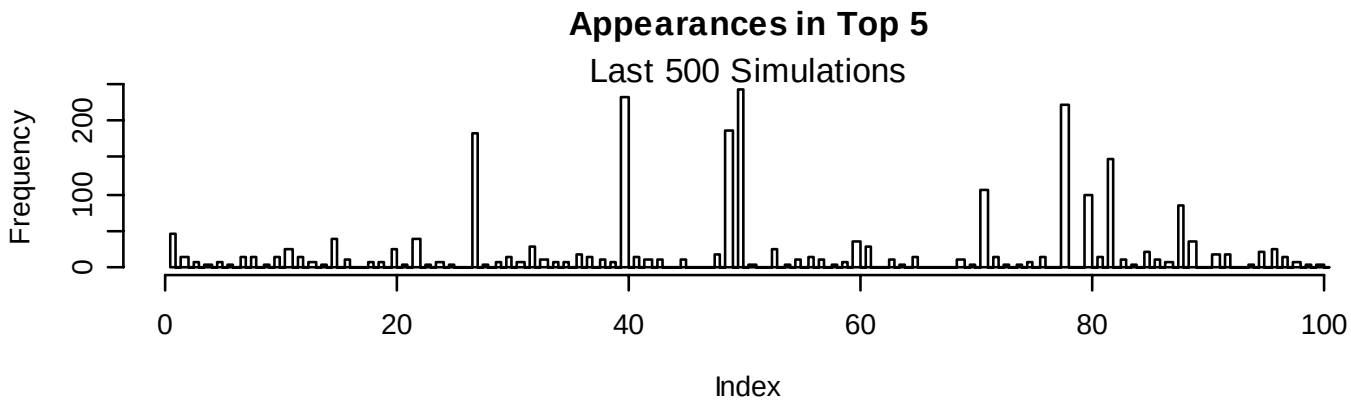
- Conclusions
  - The Accomplices plots and bubble plots both identified similar pairs of genes that appear together in “Top 5” sets most often
  - These genes tend to be most useful for dimension reduction, and hence for our classification problem
  - These Top 5 and Accomplices sets can be used to guide CCANCOR initial value selection in the event these results disagree greatly with solutions obtained from default settings.

Step 2 – Monte Carlo

- Monte Carlo Diagnostics
  - How do we know we have taken enough Monte Carlo runs to reach stable results?
  - The simcompare plot provides Top 5 distributions for the first half of runs and the last half.
    - If these are visually similar, enough runs have been taken.


Appearances in Top 5
All 1000 Simulations
Frequency
0
200
400
0
20
40
60
80
100
Index


Appearances in Top 5
First 500 Simulations
Frequency
0
100
200
0
20
40
60
80
100
Index


Appearances in Top 5
Last 500 Simulations
Frequency
0
100
200
0
20
40
60
80
100
Index