Ovarian Cancer Histotypes: Report of Statistical Findings

Derek Chiu

2021-02-14

Contents

List of Tables

List of Figures

Preface

This report of statistical findings describes the classification of ovarian cancer histotypes using data from NanoString CodeSets.

Marina Pavanello conducted the initial exploratory data analysis, Cathy Tang implemented class imbalance techniques, Derek Chiu conducted the normalization and statistical analysis, and Lauren Tindale and Aline Talhouk are the project leads.

1. Introduction

Ovarian cancer has five major histotypes: high-grade serous carcinoma (HGSC), low-grade serous carcinoma (LGSC), endometrioid carcinoma (ENOC), mucinous carcinoma (MUC), and clear cell carcinoma (CCOC). A common problem with classifying these histotypes is that there is a class imbalance issue. HGSC dominates the distribution, commonly accounting for 70% of cases in many patient cohorts, while the other four histotypes are spread over the rest of the cases.

In the NanoString CodeSets, we also run into a problem with trying to find suitable control pools to normalize the gene expression. For prospective NanoString runs, the pools can be specifically chosen, but for retrospective runs, we have to utilize a combination of common samples and common genes as references for normalization.

The supervised learning is performed under a consensus framework: we consider various classification algorithms and use evaluation metrics to help make decisions of which methods to carry forward for downstream analysis.

2. Methods

2.1 Data Processing

RNA was extracted from FFPE ovarian carcinoma samples and expression was quantified using NanoString nCounter. Samples were run in three CodeSets. Some samples or pools of samples were repeated across CodeSets for expression normalization. Normalizing CS2 to CS3 can easily follow the PrOType method for HGSC subtypes because both CodeSets have pool samples. A different technique is implemented when normalizing across CS1, CS2, and CS3 where we use common samples and genes as reference sets.

2.1.1 Raw Data

NanoString CodeSets contained a mix of all probes of interest, six positive controls spiked-in at fixed proportional concentrations (0.125- 128 fM), and eight negative controls (probes without a corresponding target). Gene targets also included 5 housekeeping genes: POLR1B, SDHA, PGK1, ACTB, RPL19. Gene selection was made from top ranked differential gene expression analysis between ovarian cancer histotypes and molecular subtypes of HGSC, as well as containing some genes of interest from unrelated projects. Gene targets in each subsequent CodeSet were re-curated, where non-informative genes were dropped and new potential differentiating genes were added.

There are 3 NanoString CodeSets:

- CS1: OvCa2103_C953
  - Samples = 412
  - Genes = 275
- CS2: PrOTYPE2_v2_C1645
  - Samples = 1223
  - Genes = 384
- CS3: OTTA2014_C2822
  - Samples = 5424
  - Genes = 532

These datasets contain raw counts extracted straight from NanoString RCC files.

2.1.2 Housekeeping Genes

The first normalization step is to normalize all endogenous genes to housekeeping genes (POLR1B, SDHA, PGK1, ACTB, RPL19; reference genes expressed in all cells). We normalize by subtracting the average $\log_2$ housekeeping gene expression from the $\log_2$ endogenous gene expression:

$\log_2$ endogenous expression - $\log_2$ average housekeeping expression = relative expression

The updated CodeSet dimensions are now:

- CS1: OvCa2103_C953
  - Samples = 412
  - Genes = 256
- CS2: PrOTYPE2_v2_C1645
  - Samples = 1223
  - Genes = 365
- CS3: OTTA2014_C2822
  - Samples = 5424
  - Genes = 513

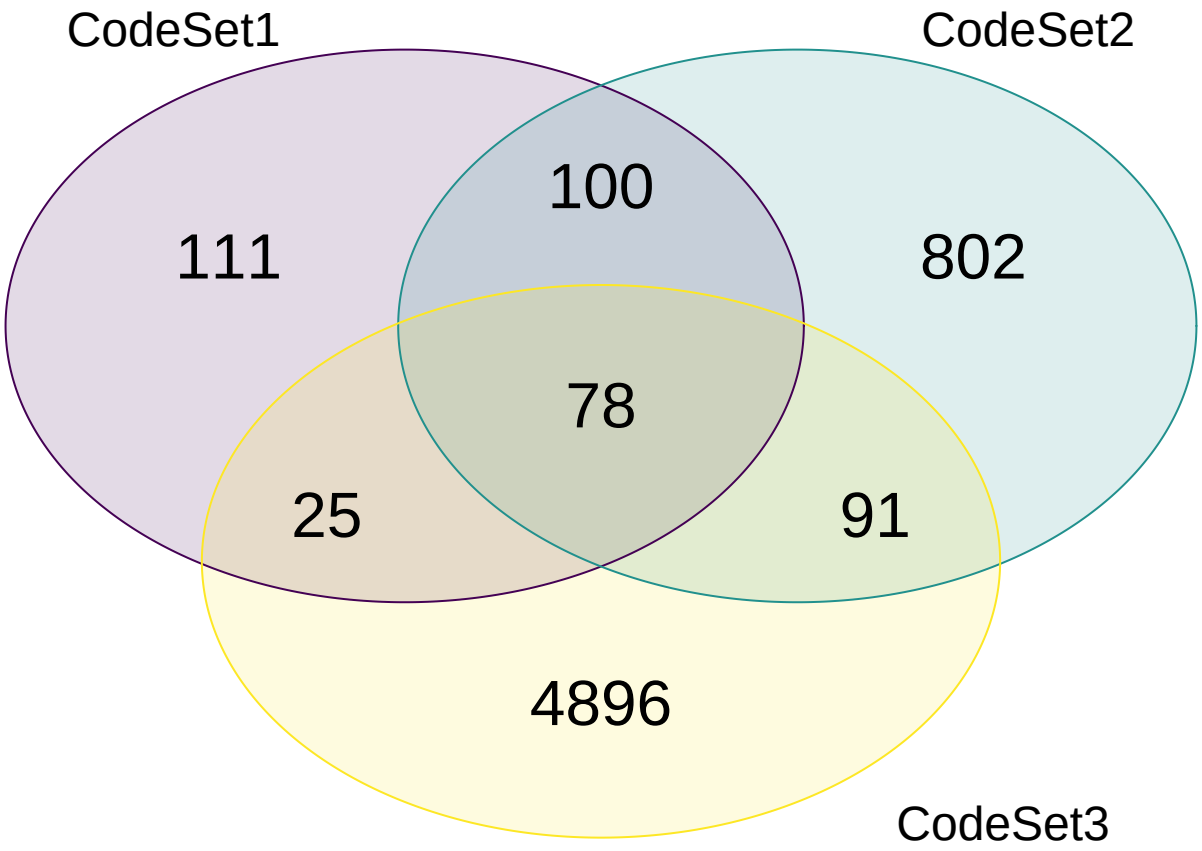
The number of genes are reduced by 19: 5 housekeeping, 8 negative, 6 positive (the latter 2 types are not used).

2.1.3 Common Samples and Genes

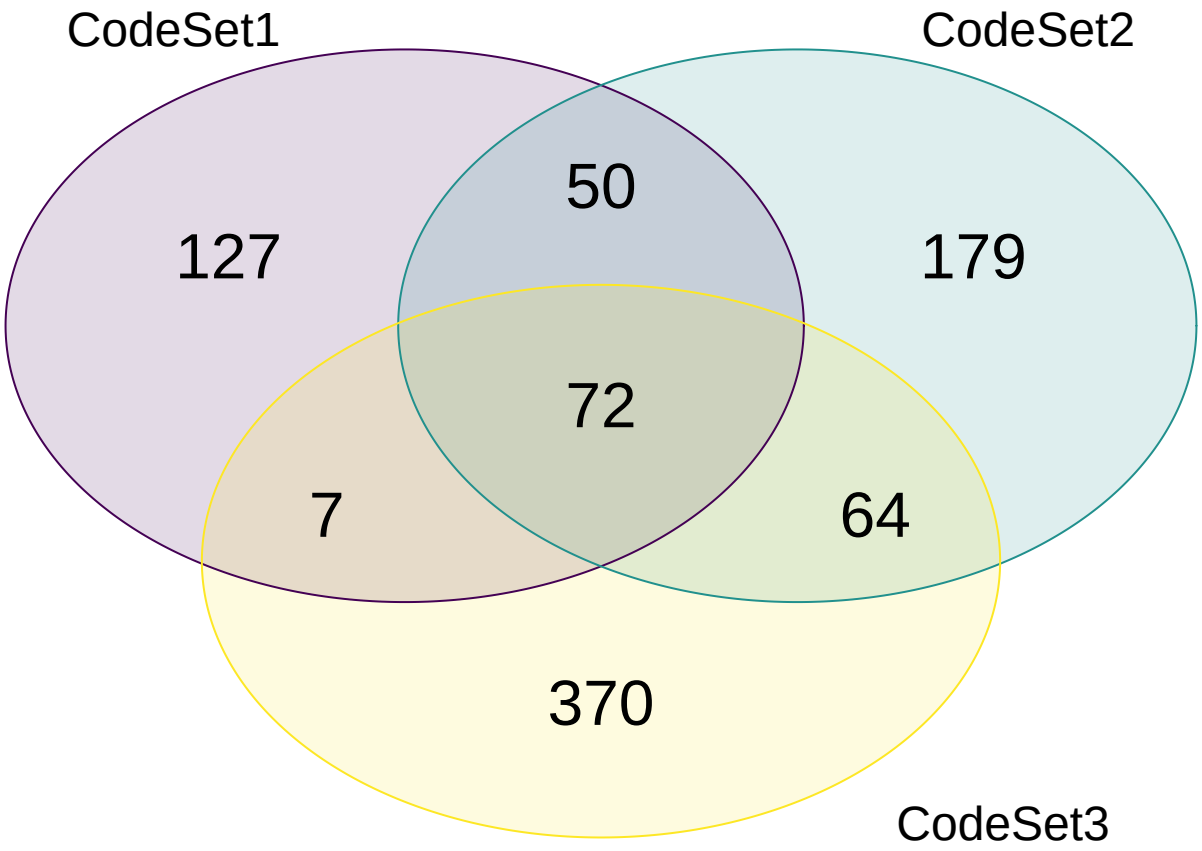
Since the reference pool samples only exist in CS2 and CS3, we need to find an alternative method to normalize all three CodeSets. One method is to select common samples and common genes that exist in all three. We found 72 common genes. Using the `summaryID` identifier, we also found 78 common summary IDs, translating to 320 samples. The number of samples that were matched to each CodeSet differed:

- CS1: OvCa2103_C953
  - Samples = 93
  - Genes = 72
- CS2: PrOTYPE2_v2_C1645
  - Samples = 87
  - Genes = 72
- CS3: OTTA2014_C2822
  - Samples = 140
  - Genes = 72

2.1.3.1 Overlap of common samples by summary ID



2.1.3.2 Overlap of common genes



*Excluding housekeeping genes and controls

2.1.4 CS1 Training Set Generation

We use the reference method to normalize CS1 to CS3.

- CS1 reference set: duplicate samples from CS1
  - Samples = 25
  - Genes = 72
- CS3 reference set: corresponding samples in CS3 also found in CS1 reference set
  - Samples = 20
  - Genes = 72
- CS1 validation set: remaining CS1 samples with reference set removed
  - Samples = 387
  - Genes = 72

The final CS1 training set has 304 samples on 72 genes after normalization and keeping only the major histotypes of interest.

Table 2.1: Cohort Distribution amongst CodeSets

cohort	cs1	cs2	cs3
MAYO	6	63	NA
MTL	3	59	NA
OOU	108	43	19
OOUE	32	30	11
VOA	145	122	538
ICON7	NA	416	NA
JAPAN	NA	8	NA
OVAR3	NA	150	NA
POOL-CTRL	NA	12	NA
DOVE4	NA	NA	1160
POOL-1	NA	NA	31
POOL-2	NA	NA	14
POOL-3	NA	NA	13
TNCO	NA	NA	691

2.1.5 CS2 Training Set Generation

We use the pool method to normalize CS2 to CS3 so we can be consistent with the PrOType normalization when there are available pools.

- CS2 pools:
  - Samples = 12 (Pool 1 = 4, Pool 2 = 4, Pool 3 = 4)
  - Genes = 365
- CS3 pools:
  - Samples = 22 (Pool 1 = 12, Pool 2 = 5, Pool 3 = 5)
  - Genes = 513
- CS2 validation set: CS2 samples with pools removed
  - Samples = 1214
  - Genes = 365

The final CS2 training set has 945 samples on 136 (common) genes after normalization and keeping only the major histotypes of interest.

2.1.6 Cohort Distribution

CodeSets comprised samples from sites collected internationally as shown below. Note that the CS3 pools sample total (n=58) shown here include those that are not used as reference pools, following previous normalization methods. In particular, the distribution of CS3 pools actually used for normalization (n=22) is POOL1 = 12, POOL2 = 5, POOL3 = 5.

2.2 Normalization Between CodeSets

After normalization to housekeeping genes and filtering for the five major histotypes of interest, as determined by pathology review and/or IHC, two methods were used to normalize data between CodeSets.

Table 2.2: Distinct Cohort Distribution amongst CodeSets

cohort	cs1	cs2	cs3
MAYO	6	62	NA
MTL	3	59	NA
OOU	99	43	19
OOUE	31	30	11
VOA	136	107	452
ICON7	NA	383	NA
JAPAN	NA	8	NA
OVAR3	NA	150	NA
POOL-CTRL	NA	3	NA
DOVE4	NA	NA	1094
POOL-1	NA	NA	12
POOL-2	NA	NA	5
POOL-3	NA	NA	5
TNCO	NA	NA	674

2.2.1 Common Samples Method

The common samples method was used to normalize CodeSet1, 2, and 3, where common samples and genes were used as reference sets. Among the samples repeated in all CodeSets we normalized using either: a random set of 3 samples from each major histotype (random3; n=15), a random set of 2 samples from each major histotype (random2; n=10), or a random set of 1 sample from each major histotype (random1; n=5). In each case CodeSet3 expression (X_3) was held fixed, while CodeSet1/2 expression (X_1 and X_2) were normalized to CodeSet3 by subtracting the average gene expression from the CodeSet1/2 reference set (R_1 or R_2) and adding the average gene expression of the CodeSet3 reference set (R_3). Alternatively, X_1 (norm) = X_1 - R_1 + R_3 would calibrate CodeSet1 to CodeSet3.

2.2.2 Pools Method

The pools method was used to normalize CodeSet2 and CodeSet3. The three reference pools, regularly assayed mixes of samples representing all histotypes, were run in CodeSet2 and CodeSet3 only. CodeSet2 contained 12 reference pool samples (Pool 1 = 4, Pool 2 = 4, Pool 3 = 4) and CodeSet3 contained 22 reference pool samples (Pool 1 = 12, Pool 2 = 5, Pool 3 = 5). Similar to the common samples method, CodeSet2 was normalized to CodeSet3 via: X_2 (norm) = X_2 - R_2 + R_3 where R is the average expression of the reference pool samples in the respective CodeSet. This method of pool normalization was also used by PrOType to classify HGSC subtypes

2.2.3 Concordance Comparison

Concordance between CodeSets using the different normalization strategies was compared in common samples, excluding those used for the normalization, using Pearson's correlation coefficient (R^2), coefficient of accuracy (Ca), and Lin's concordance correlation (Rc = R^2 x Ca).

2.3 Histotype Classification

We use 5 classification algorithms and 4 subsampling methods across 500 repetitions in the supervised learning framework for the Training Set, CS1 and CS2. The pipeline was run using SLURM batch jobs

submitted to a partition on a CentOS 7 server. Implementations of the techniques below were called from the splendid package.

- Classifiers:
  - Random Forest
  - SVM
  - Adaboost
  - Multinomial Regression Model with Ridge Penalty
  - Multinomial Regression Model with LASSO Penalty
- Subsampling:
  - None
  - Down-sampling
  - Up-sampling
  - SMOTE

3. Validation

3.1 Full Data Distributions

The histotype distributions on the full data are shown below.

3.2 Training Set Distributions

The training set distributions for CS1 and CS2 are shown below.

Table 3.1: All CodeSet Histotype Groups

hist_gr	CS1	CS2	CS3
HGSC	169	757	2453
non-HGSC	196	373	677

Table 3.2: All CodeSet Histotypes

revHist	CS1	CS2	CS3
CARCINOMA-NOS	0	61	23
Carcinoma, NOS	0	0	2
CCOC	57	68	182
CCOC-MCT	0	1	0
Cell-Line	17	48	13
CTRL	0	12	0
ENOC	61	30	272
ENOC-CCOC	0	7	0
ERROR	0	3	0
HGSC	169	757	2453
HGSC-MCT	0	1	0
LGSC	22	29	50
MBOT	0	20	3
MET-NOP	0	21	0
MIXED (ENOC/CCOC)	0	0	1
MIXED (ENOC/LGSC)	0	0	1
MIXED (HGSC/CCOC)	0	0	1
mixed cell	0	0	7
MMMT	0	0	30
MUC	20	61	77
Other (use when 6, 7, or 9 is not distinguished) or unknown if epithelial	0	0	1
Other/Exclude	0	0	8
SBOT	19	10	3
Serous	0	0	2
serous LMP	0	0	1
SQAMOUS	0	1	0

Table 3.3: Common Summary ID CodeSet Histotypes

revHist	CS1	CS2	CS3
CCOC	3	4	9
Cell-Line	4	5	5
ENOC	4	4	9
HGSC	68	64	98
LGSC	7	5	8
MUC	7	5	11

Table 3.4: All CodeSet Major Histotypes

revHist	CS1	CS2	CS3	CS1_percent	CS2_percent	CS3_percent
CCOC	57	68	182	17.3	7.2	6.0
ENOC	61	30	272	18.5	3.2	9.0
HGSC	169	757	2453	51.4	80.1	80.9
LGSC	22	29	50	6.7	3.1	1.6
MUC	20	61	77	6.1	6.5	2.5

Table 3.5: CS1 Histotypes

CodeSet	revHist	n
CS1	CCOC	57
CS1	Cell-Line	17
CS1	ENOC	61
CS1	HGSC	169
CS1	LGSC	22
CS1	MUC	20
CS1	SBOT	19

Table 3.6: CS2 Histotypes

CodeSet	revHist	n
CS2	CARCINOMA-NOS	61
CS2	CCOC	68
CS2	CCOC-MCT	1
CS2	Cell-Line	48
CS2	CTRL	12
CS2	ENOC	30
CS2	ENOC-CCOC	7
CS2	ERROR	3
CS2	HGSC	757
CS2	HGSC-MCT	1
CS2	LGSC	29
CS2	MBOT	20
CS2	MET-NOP	21
CS2	MUC	61
CS2	SBOT	10
CS2	SQAMOUS	1

Table 3.7: CS3 Histotypes

CodeSet	revHist	n
CS3	CARCINOMA-NOS	23
CS3	Carcinoma, NOS	2
CS3	CCOC	182
CS3	Cell-Line	13
CS3	ENOC	272
CS3	HGSC	2453
CS3	LGSC	50
CS3	MBOT	3
CS3	MIXED (ENOC/CCOC)	1
CS3	MIXED (ENOC/LGSC)	1
CS3	MIXED (HGSC/CCOC)	1
CS3	mixed cell	7
CS3	MMMT	30
CS3	MUC	77
CS3	Other (use when 6, 7, or 9 is not distinguished) or unknown if epithelial	1
CS3	Other/Exclude	8
CS3	SBOT	3
CS3	Serous	2
CS3	serous LMP	1

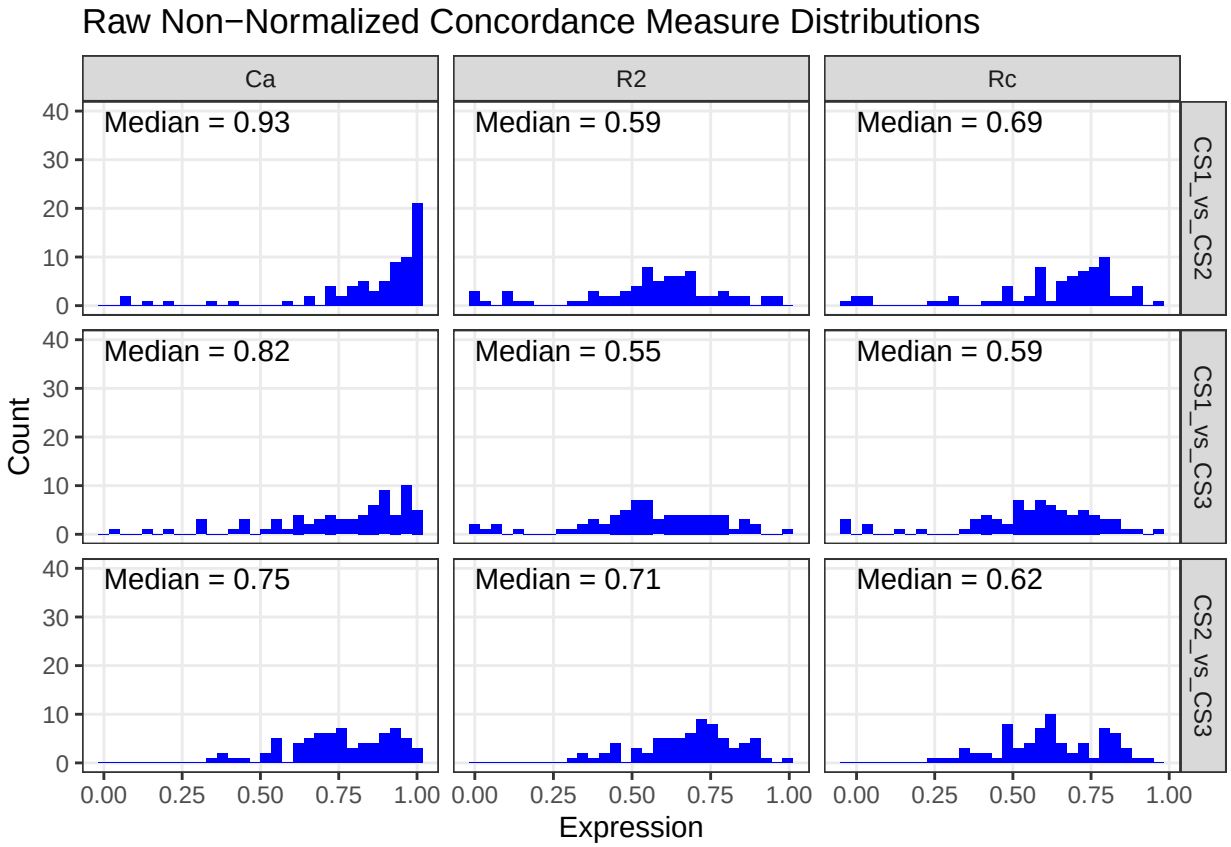
Table 3.8: CS1 Training Set Histotypes

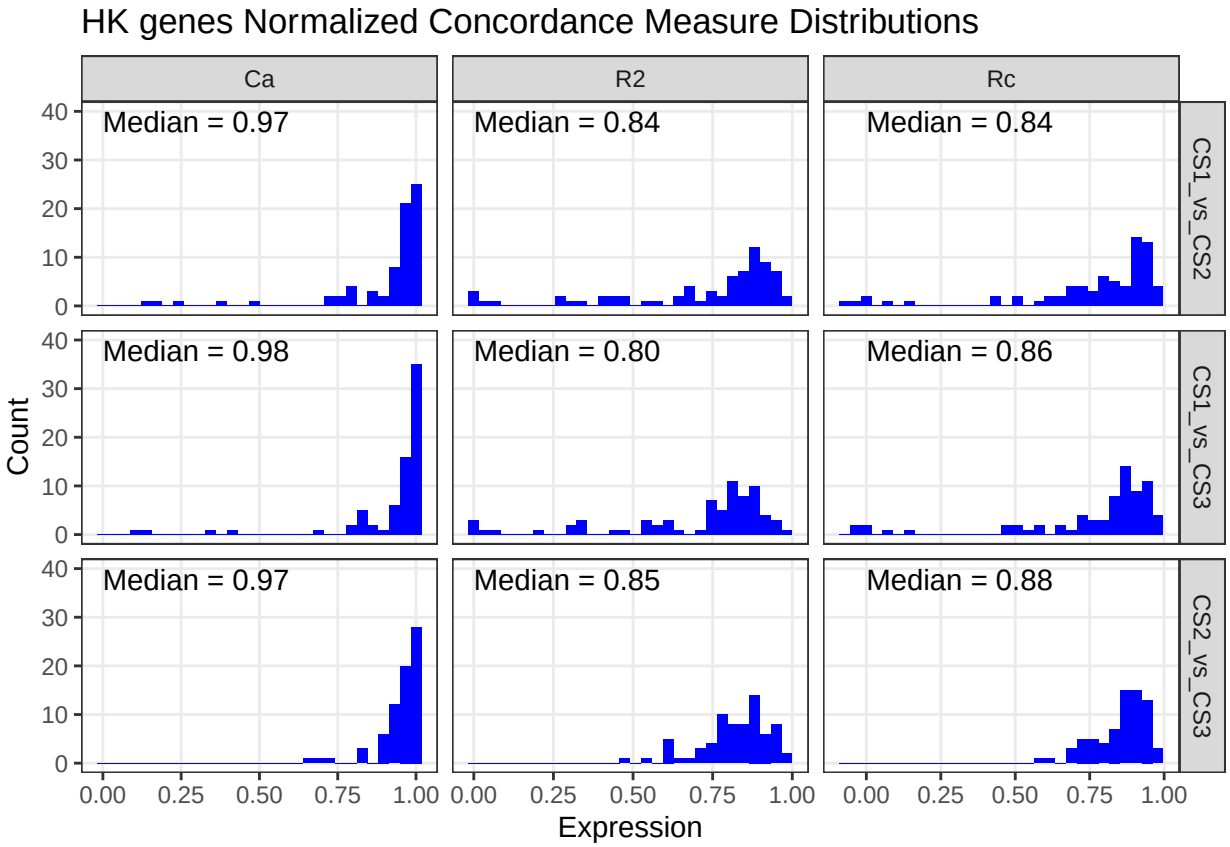
histotype	n
CCC	57
ENOCa	59
HGSC	156
LGSC	16
MUC	16

Table 3.9: CS2 Training Set Histotypes

histotype	n
CCOC	68
ENOC	30
HGSC	757
LGSC	29
MUC	61

3.3 Normalization





3.3.1 Common Samples Method

We employ a new normalization technique using randomly selected samples common to all three CodeSets with a uniform distribution of histotypes as the reference dataset. The number of randomly selected samples ranges from 1-3 per histotype. Hence, the reference dataset has either 5, 10, or 15 samples and we validate on the remaining. Note that ottaID duplicates are collapsed by mean averaging the gene expression. There are n=72 common samples.

CodeSets 1 and 2 are calibrated to CodeSet3 as follows:

- `X^1(norm) = X^1 - R^1 + R^3`
- `X^2(norm) = X^2 - R^2 + R^3`
- `X^3(norm) = X^3`

3.3.1.1 Random3

Randomly choose 3 samples from each of the 5 histotypes as the reference set (n=15). The rest are validated.

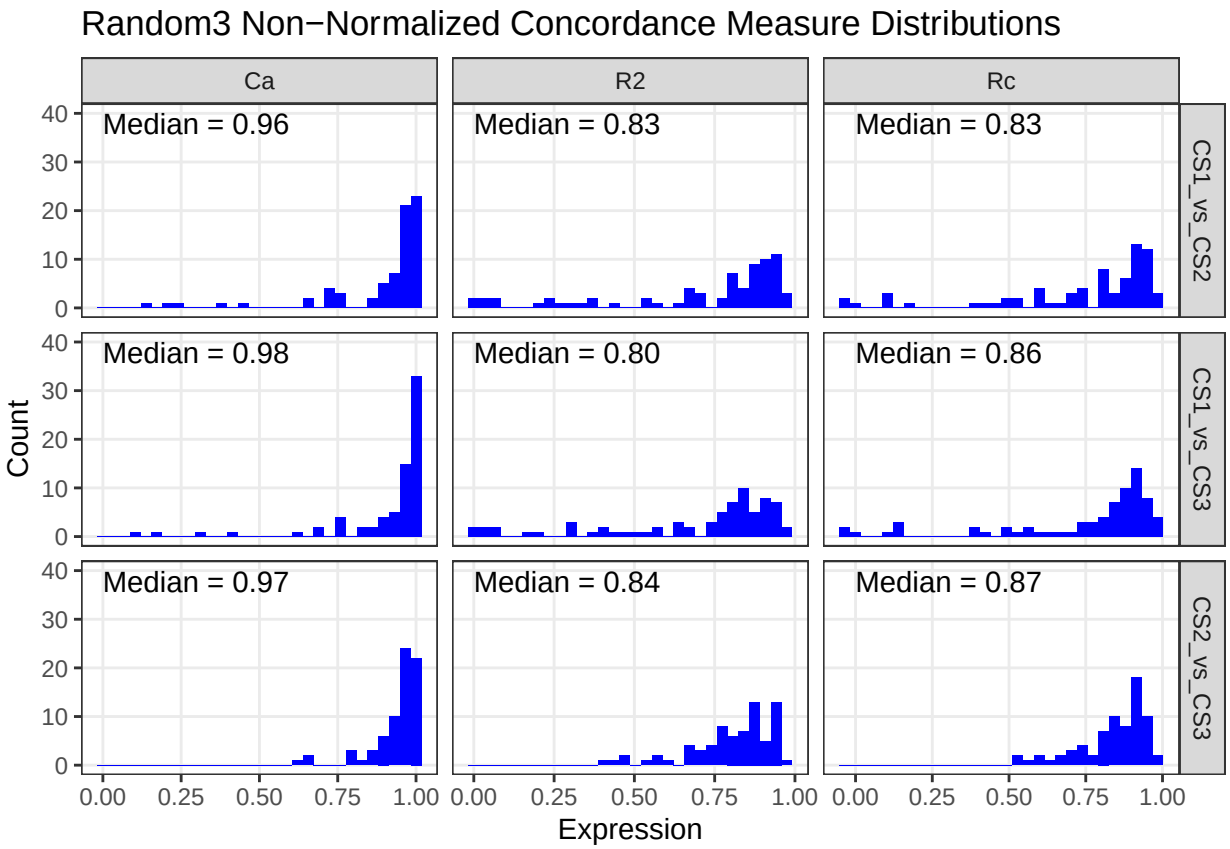


Figure 3.1: Random3 Non-Normalized Concordance Measure Distributions

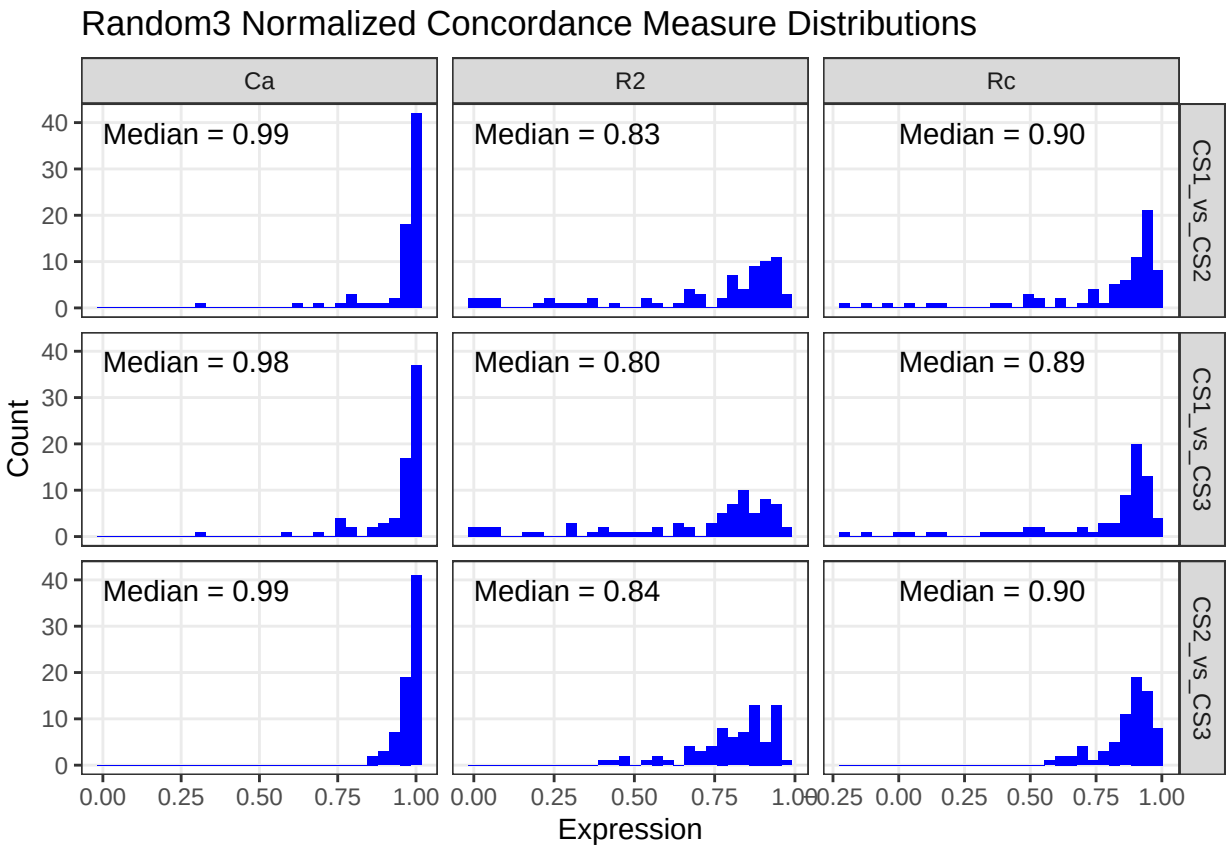


Figure 3.2: Random3 Normalized Concordance Measure Distributions

3.3.1.2 Random2

Randomly choose 2 samples from each of the 5 histotypes as the reference set (n=10). The rest are validated.

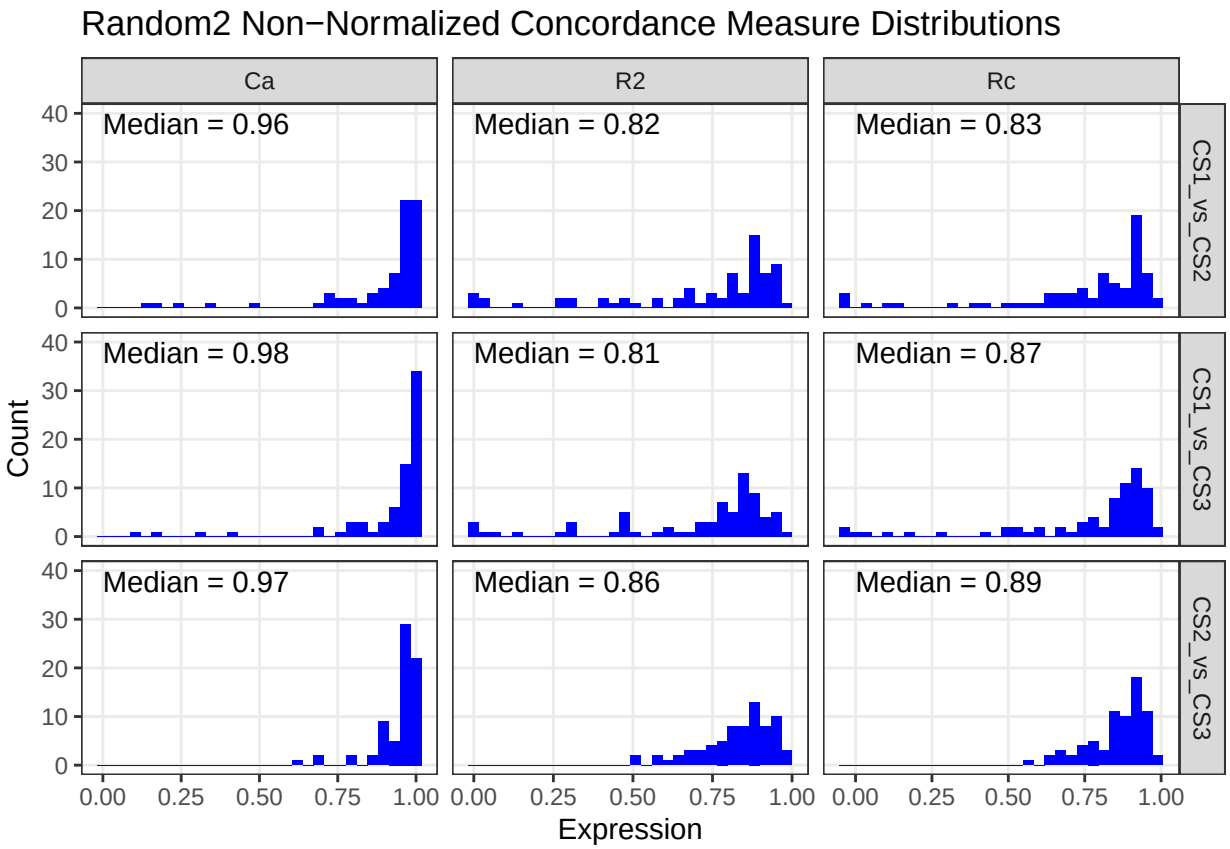


Figure 3.3: Random2 Non-Normalized Concordance Measure Distributions

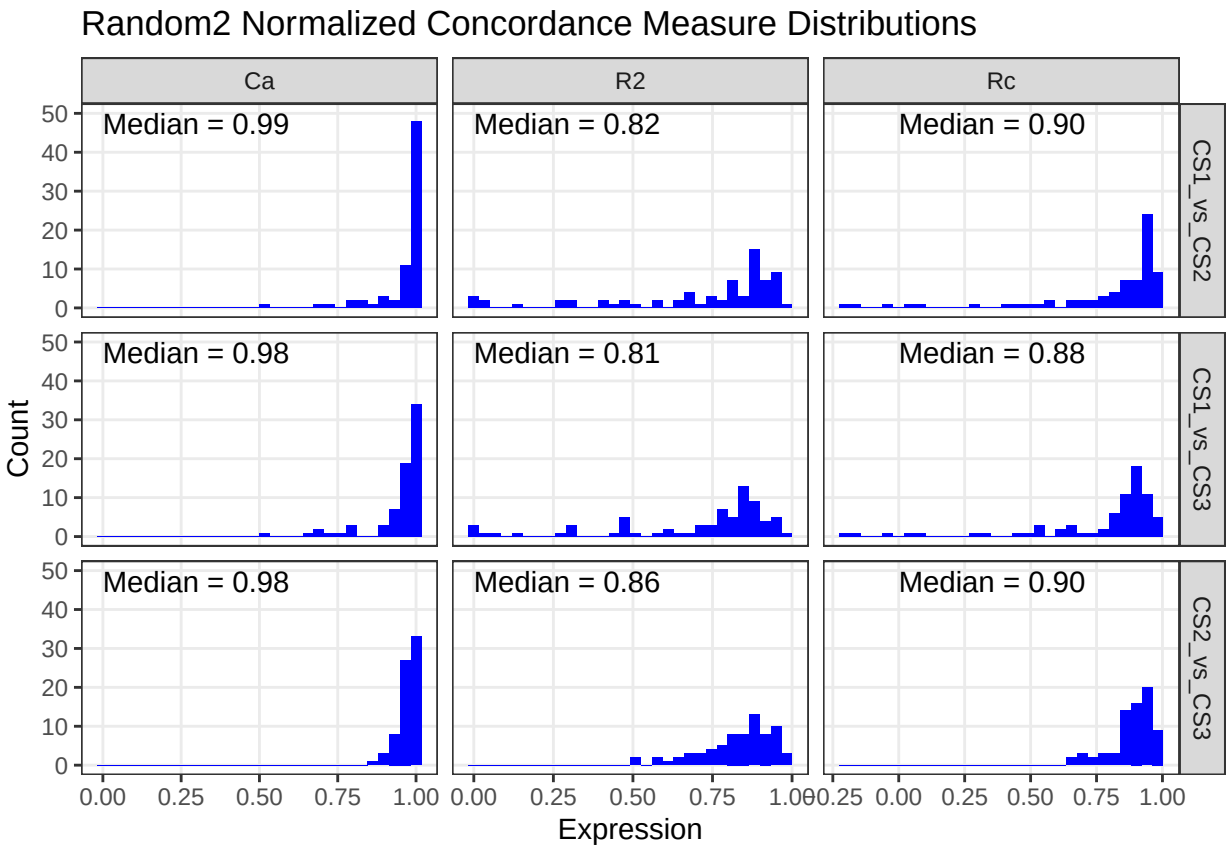


Figure 3.4: Random2 Normalized Concordance Measure Distributions

3.3.1.3 Random1

Randomly choose 1 sample from each of the 5 histotypes as the reference set (n=5). The rest are validated.

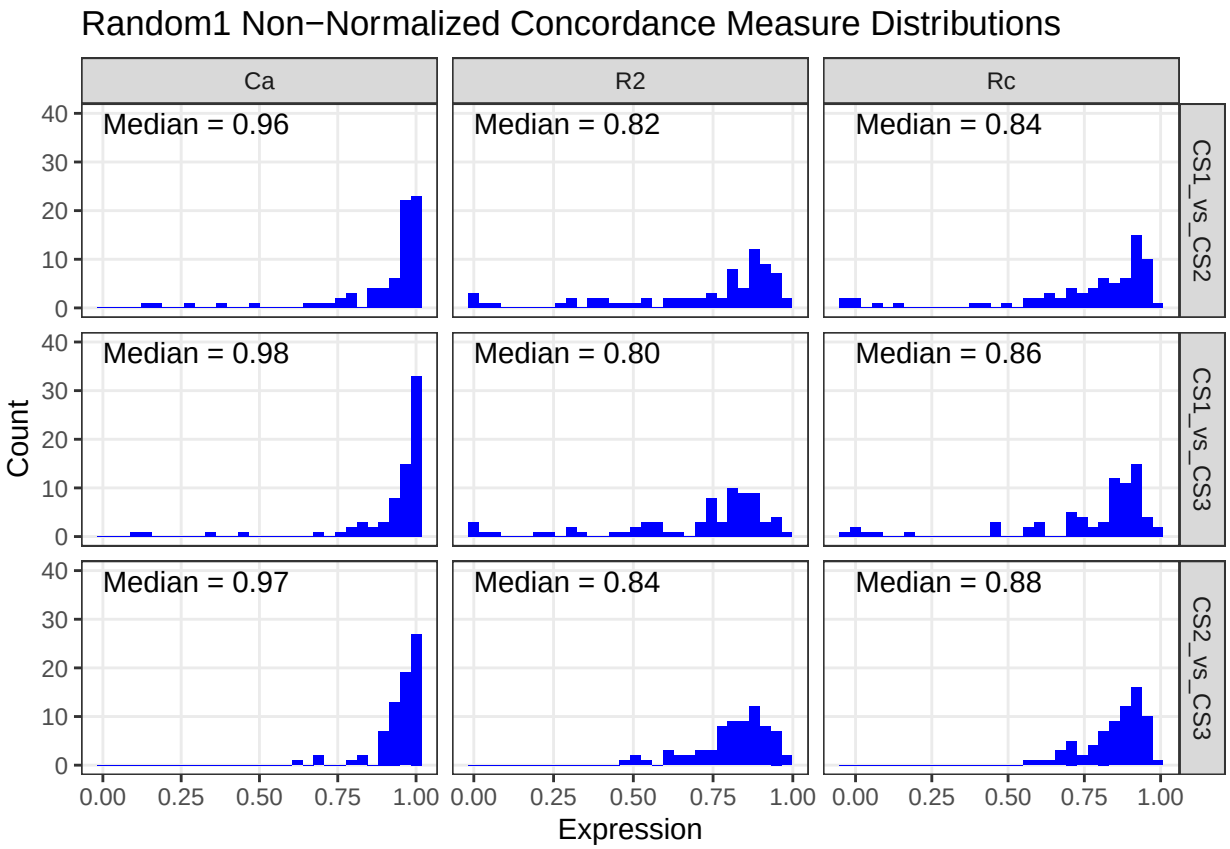


Figure 3.5: Random1 Non-Normalized Concordance Measure Distributions

Table 3.10: Random1 CS1 vs. CS3 Median Concordance Measures by Histotypes

hist	R2-Non	Ca-Non	Rc-Non	R2-Norm	Ca-Norm	Rc-Norm
CCOC	1.00	0.29	0.12	1.00	0.29	0.10
ENOC	1.00	0.54	0.54	1.00	0.62	0.62
HGSC	0.79	0.98	0.85	0.79	0.97	0.87
LGSC	0.96	0.89	0.82	0.96	0.91	0.87
MUC	0.77	0.86	0.68	0.77	0.81	0.63

Table 3.11: Random1 CS2 vs. CS3 Median Concordance Measures by Histotypes

hist	R2-Non	Ca-Non	Rc-Non	R2-Norm	Ca-Norm	Rc-Norm
CCOC	1.00	0.23	0.08	1.00	0.27	0.16
ENOC	1.00	0.63	0.61	1.00	0.61	0.57
HGSC	0.83	0.96	0.86	0.83	0.98	0.89
LGSC	0.98	0.92	0.90	0.98	0.95	0.93
MUC	0.68	0.77	0.55	0.68	0.86	0.61

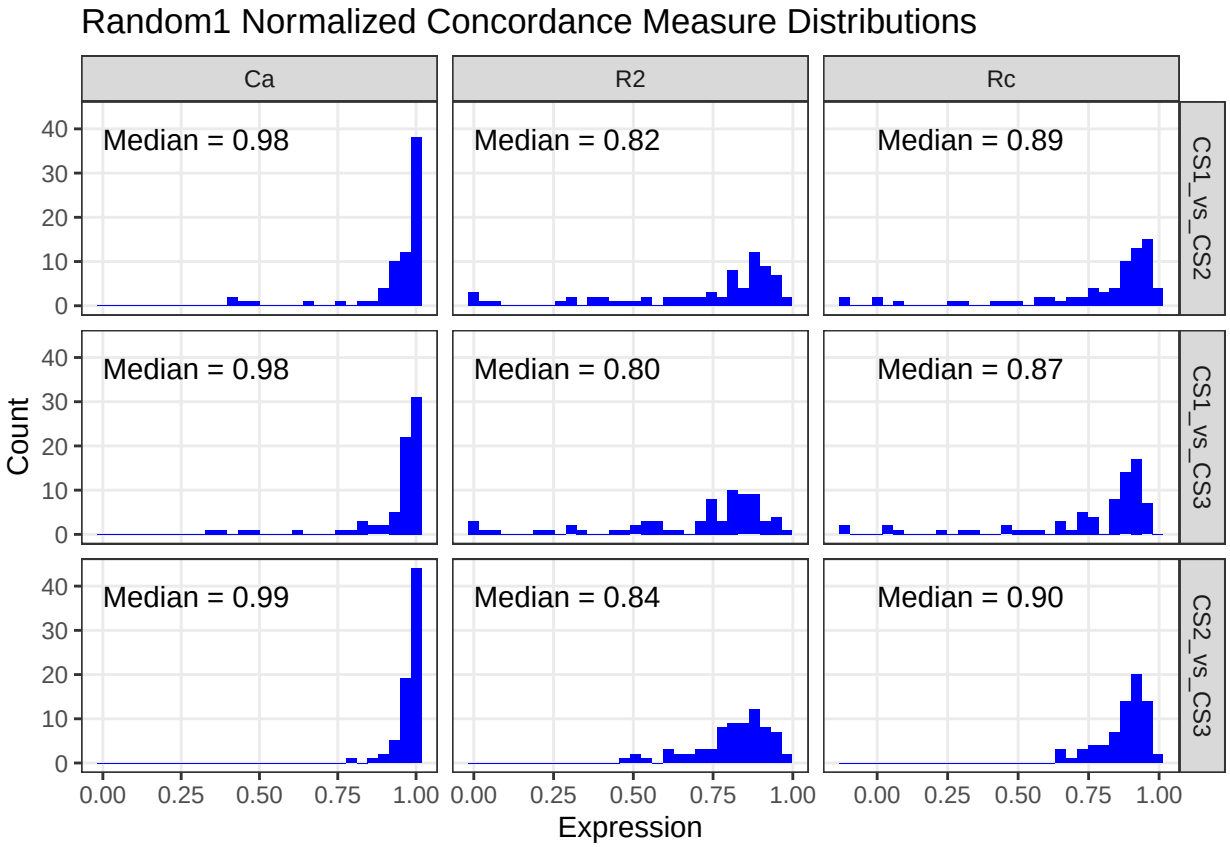


Figure 3.6: Random1 Normalized Concordance Measure Distributions

In Tables 3.10 and 3.11, we calculate the concordance measures for CS1 vs. CS3 and CS2 cs. CS3, respectively. The measures are calculated for both non-normalized and normalized datasets (CS1, CS2), and split by histotype.

Table 3.12: Random3 HGSC CS1 vs. CS3 Median Concordance Measures by Histotypes

hist	R2-Non	Ca-Non	Rc-Non	R2-Norm	Ca-Norm	Rc-Norm
CCOC	0.62	0.62	0.32	0.62	0.68	0.27
ENOC	0.88	0.76	0.66	0.88	0.77	0.70
HGSC	0.77	0.97	0.85	0.77	0.99	0.87
LGSC	0.94	0.85	0.80	0.94	0.90	0.84
MUC	0.74	0.92	0.72	0.74	0.93	0.78

3.3.1.4 Random3 HGSC

Randomly choose n=3 HGSC samples as the reference set, and use the rest as validation. This was tried in lieu of the fact that some non-HGSC histotypes have at most n=3 samples in total, so using Random3 or even Random2 would leave no samples remaining in the validation set for these histotypes.

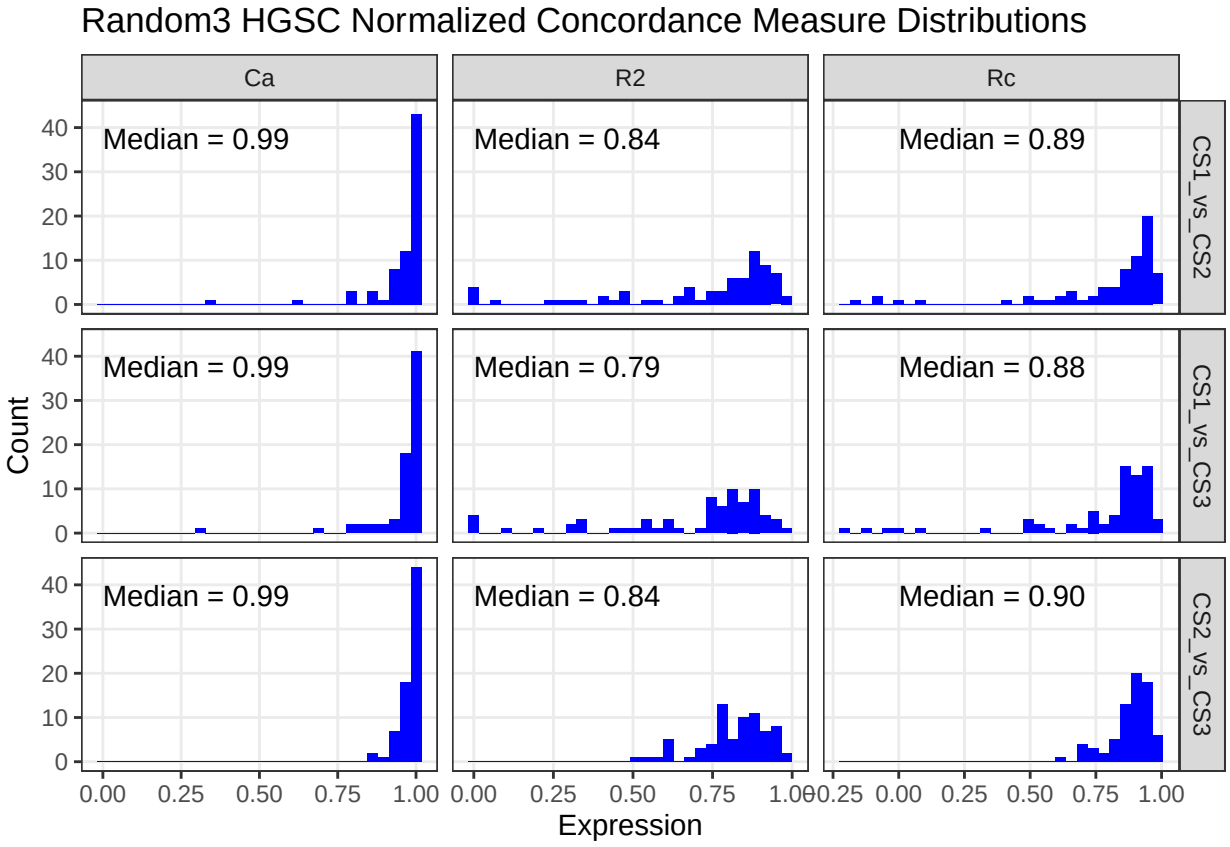


Figure 3.7: Random3 HGSC Normalized Concordance Measure Distributions

In Tables 3.12 and 3.13, we calculate the concordance measures for CS1 vs. CS3 and CS2 cs. CS3, respectively. The measures are calculated for both non-normalized and normalized datasets (CS1, CS2), and split by histotype.

Table 3.13: Random3 HGSC CS2 vs. CS3 Median Concordance Measures by Histotypes

hist	R2-Non	Ca-Non	Rc-Non	R2-Norm	Ca-Norm	Rc-Norm
CCOC	0.66	0.56	0.35	0.66	0.59	0.42
ENOC	0.85	0.76	0.66	0.85	0.85	0.76
HGSC	0.82	0.96	0.86	0.82	0.99	0.90
LGSC	0.97	0.95	0.92	0.97	0.92	0.90
MUC	0.74	0.89	0.72	0.74	0.93	0.72

3.3.1.5 Random1 for Sites

We use the Random1 method to normalize CS3-USC and CS3-AOC to CS3-VAN. There aren't enough samples in the USC and AOC cohorts to perform Random2 or Random3.

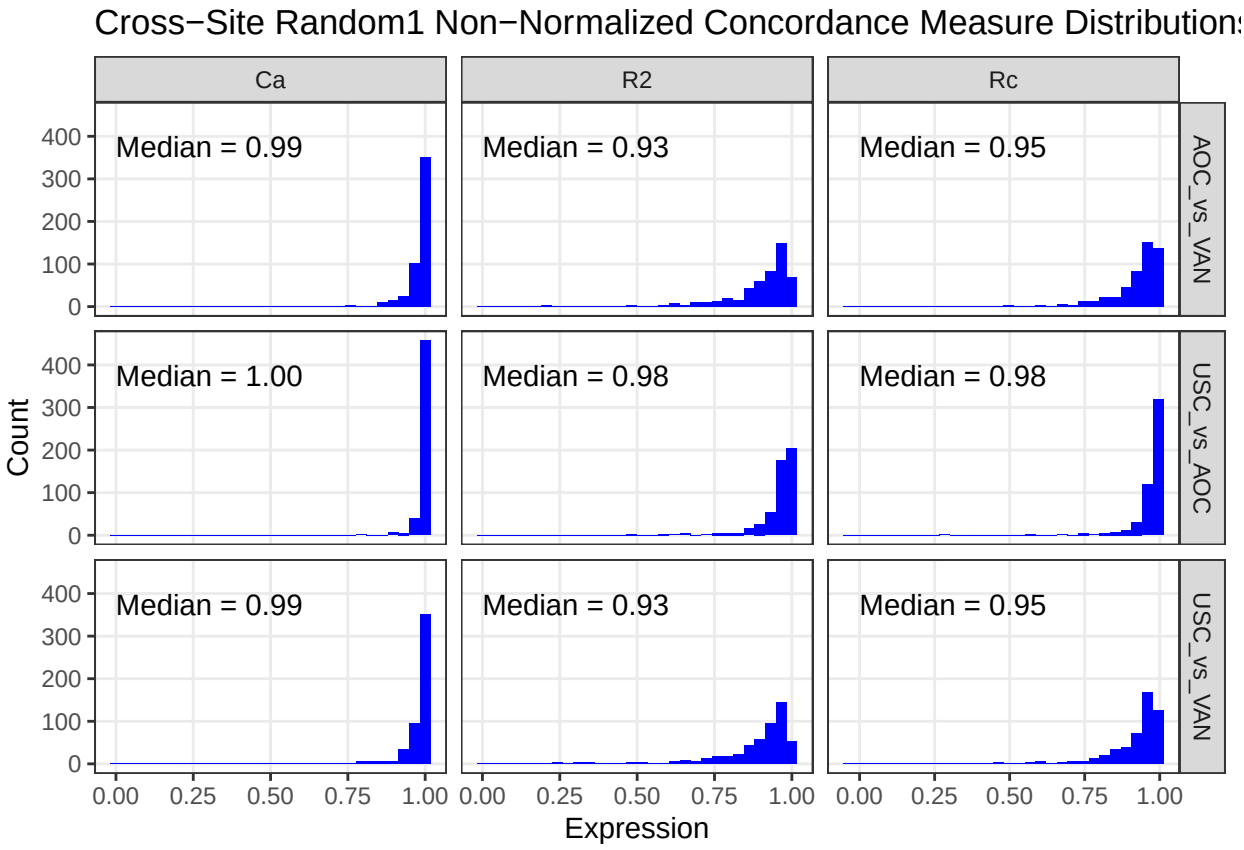
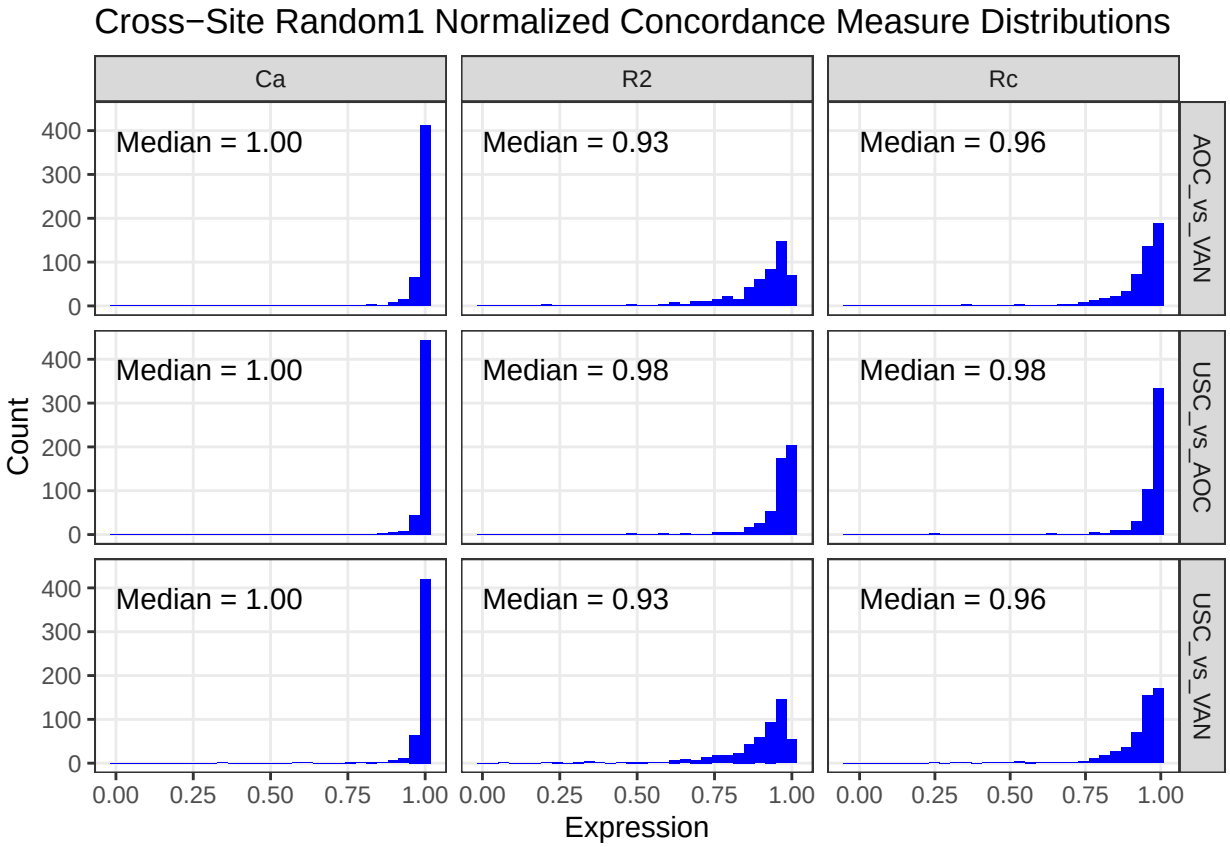


Figure 3.8: Cross-Site Random1 Non-Normalized Concordance Measure Distributions



3.3.2 Pools Method

3.3.2.1 CS2 vs. CS3

CodeSet2 contains 12 ref pool samples (Pool 1 = 4, Pool 2 = 4, Pool 3 = 4). CodeSet3 contains 22 ref pool samples (Pool 1 = 12, Pool 2 = 5, Pool 3 = 5). n=84 common samples.

CodeSet2 is calibrated to CodeSet3 as follows:

```
X^2(norm) = X^2 - R^2 + R^3
X^3(norm) = X^3
```

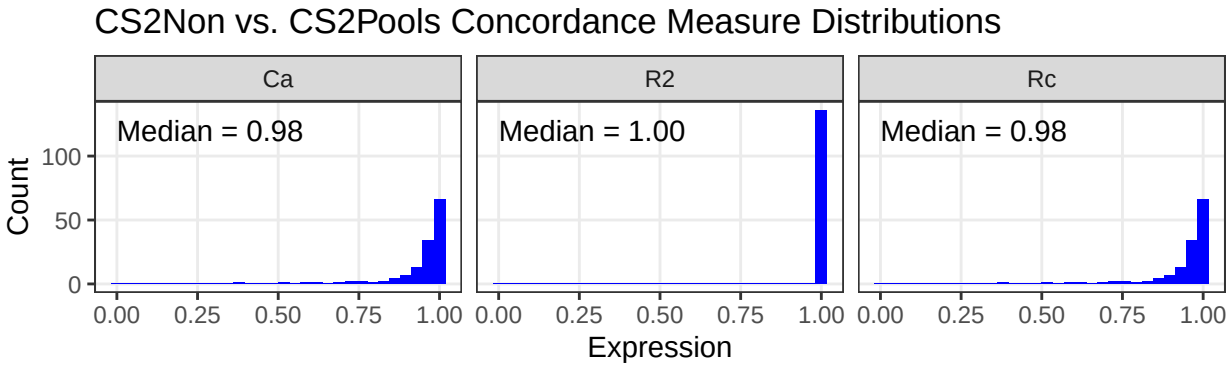


Figure 3.9: CS2Non vs. CS2Pools Concordance Measure Distributions

Table 3.14: Pools Non-Normalized CS2 vs. CS3 Median Concordance Measures by Histotypes

hist	R2	Ca	Rc
CCOC	0.66	0.53	0.26
ENOC	0.88	0.74	0.63
HGSC	0.77	0.94	0.80
LGSC	0.98	0.95	0.92
MUC	0.74	0.86	0.68

Table 3.15: Pools Normalized CS2 vs. CS3 Median Concordance Measures by Histotypes

hist	R2	Ca	Rc
CCOC	0.66	0.60	0.32
ENOC	0.88	0.76	0.68
HGSC	0.77	0.94	0.81
LGSC	0.98	0.95	0.93
MUC	0.74	0.91	0.71

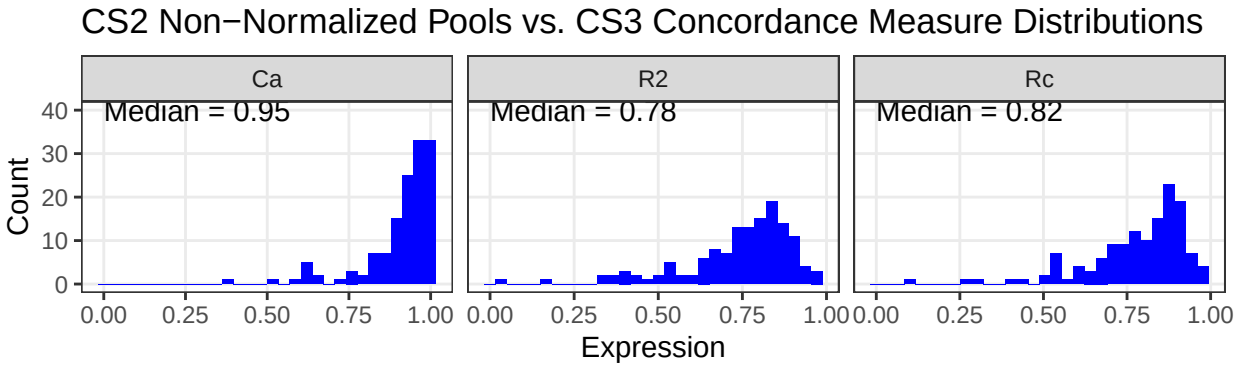


Figure 3.10: CS2 Non-Normalized Pools vs. CS3 Concordance Measure Distributions

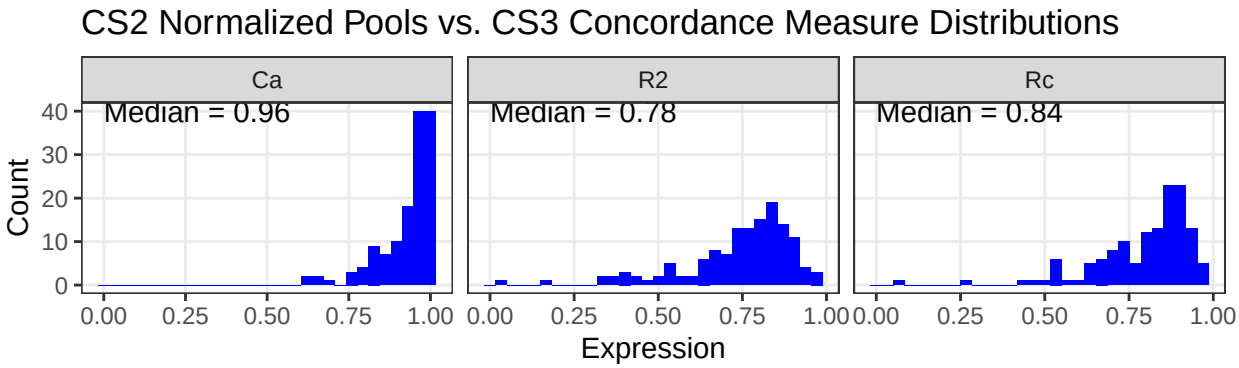


Figure 3.11: CS2 Normalized Pools vs. CS3 Concordance Measure Distributions

3.3.2.2 USC vs. VAN

In CodeSet 3, we normalize the USC and AOC cohorts to the VAN cohort which is used as the reference dataset.

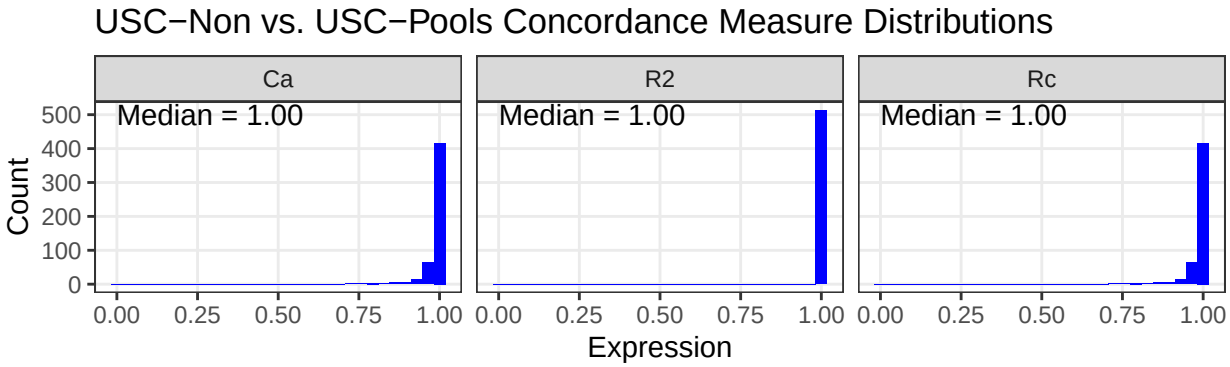


Figure 3.12: USC-Non vs. USC-Pools Concordance Measure Distributions

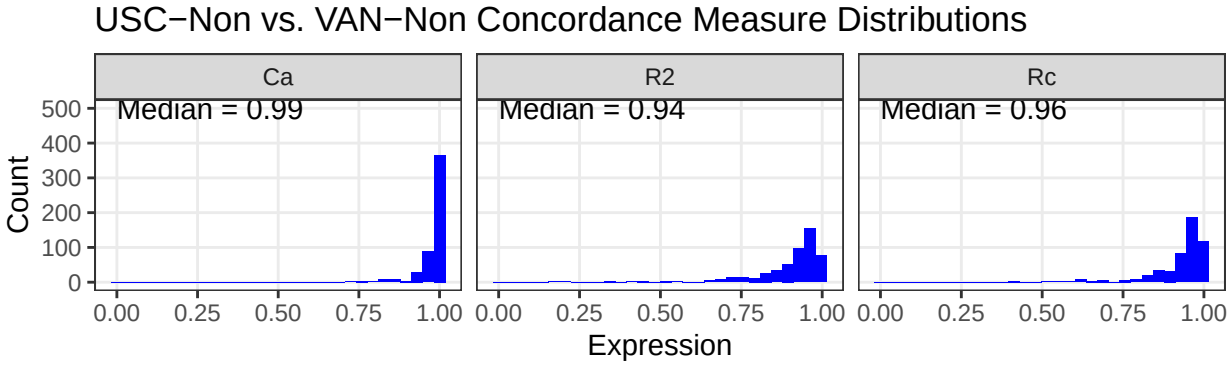


Figure 3.13: USC-Non vs. VAN-Non Concordance Measure Distributions

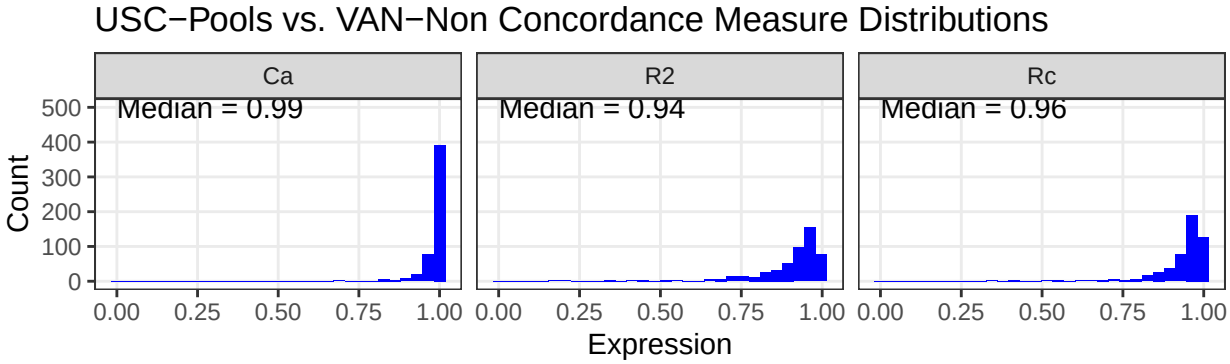


Figure 3.14: USC-Pools vs. VAN-Non Concordance Measure Distributions

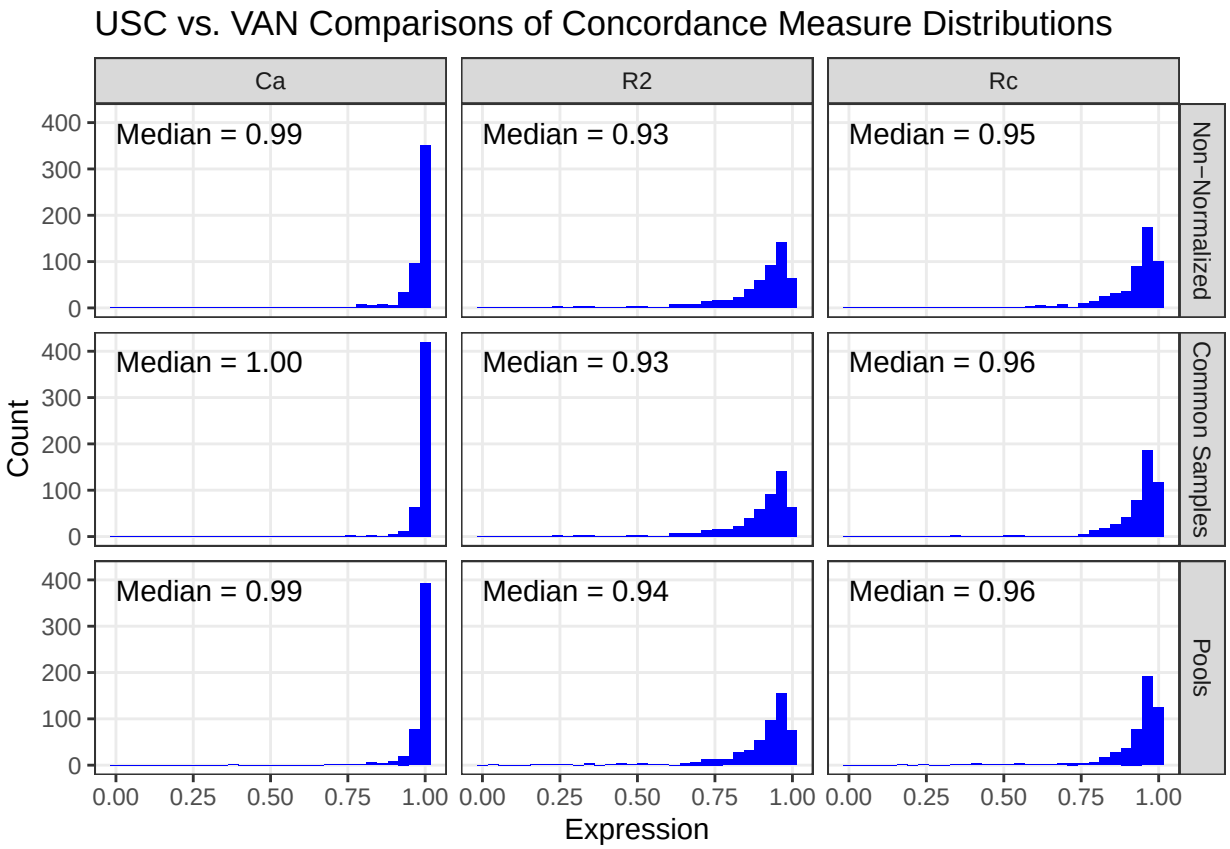


Figure 3.15: USC vs. VAN Comparisons of Concordance Measure Distributions

3.3.2.3 AOC vs. VAN

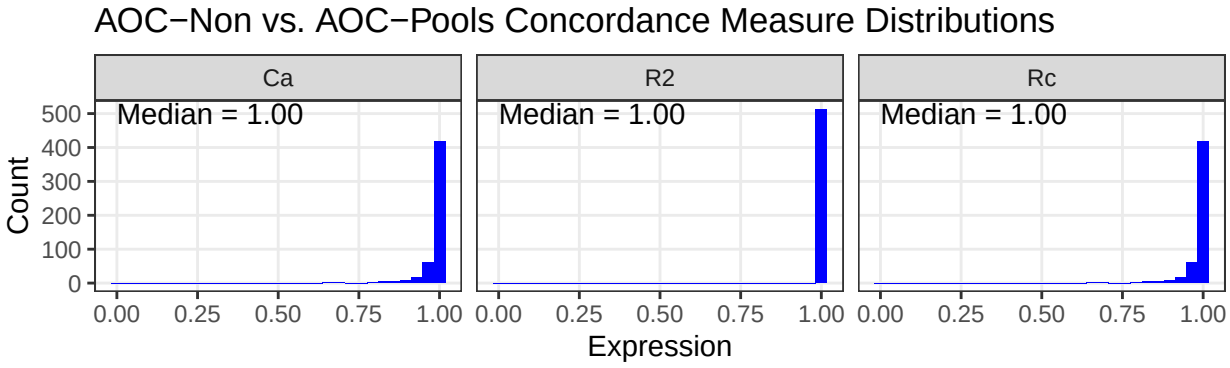


Figure 3.16: AOC-Non vs. AOC-Pools Concordance Measure Distributions

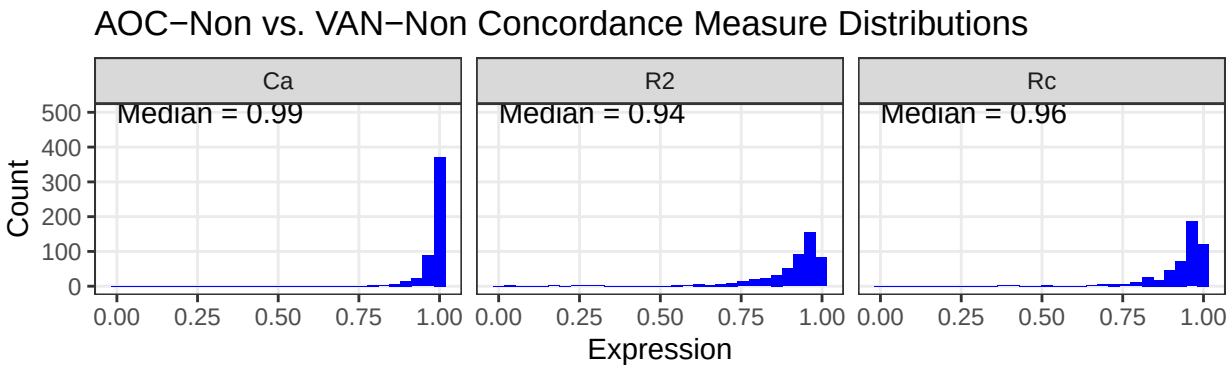


Figure 3.17: AOC-Non vs. VAN-Non Concordance Measure Distributions

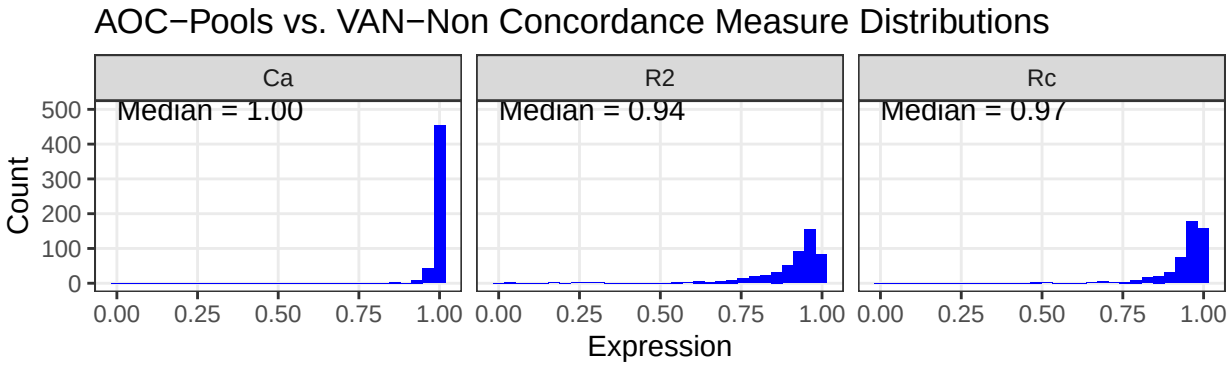


Figure 3.18: AOC-Pools vs. VAN-Non Concordance Measure Distributions

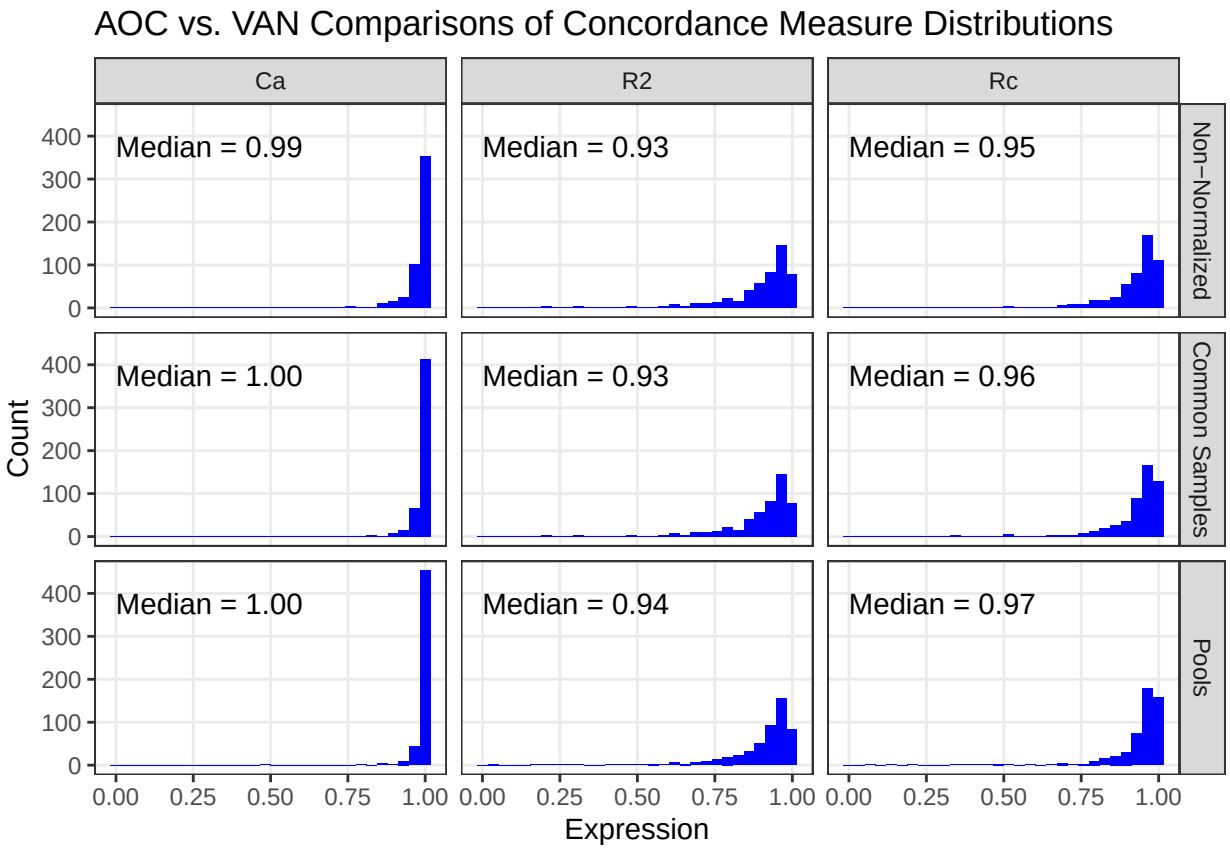


Figure 3.19: AOC vs. VAN Comparisons of Concordance Measure Distributions

3.3.3 Common Samples vs. Pools Comparison

Since only CS2 and CS3 have pools, we make three comparisons between these two CodeSets:

- Non-Normalized
- Common Samples Method
- Pools Method

Table 3.16: Random3 Samples Comparisons Statistics by Histotypes

hist	R2-Non	Ca-Non	Rc-Non	R2-Common	Ca-Common	Rc-Common	R2-Pools	Ca-Pools	Rc-Pools
HGSC	0.84	0.96	0.86	0.84	0.99	0.90	0.84	0.96	0.86
LGSC	NA	NA	NA	NA	NA	NA	NA	NA	NA
MUC	1.00	0.49	0.44	1.00	0.62	0.52	1.00	0.46	0.42

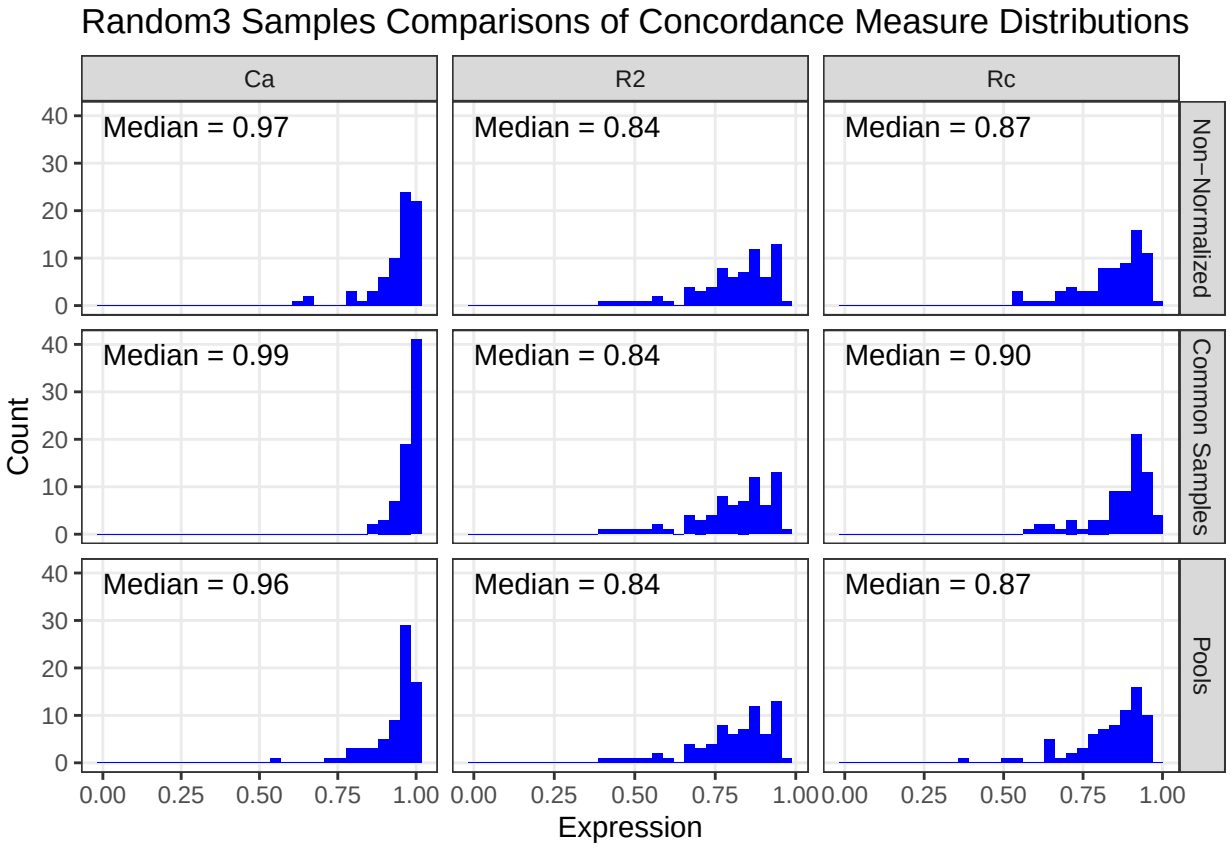


Figure 3.20: Random3 Samples Comparisons of Concordance Measure Distributions

Table 3.17: Random2 Samples Comparisons Statistics by Histotypes

hist	R2-Non	Ca-Non	Rc-Non	R2-Common	Ca-Common	Rc-Common	R2-Pools	Ca-Pools	Rc-Pools
CCOC	NA	NA	NA	NA	NA	NA	NA	NA	NA
ENOC	NA	NA	NA	NA	NA	NA	NA	NA	NA
HGSC	0.84	0.96	0.87	0.84	0.98	0.89	0.84	0.96	0.86
LGSC	1.00	0.88	0.87	1.00	0.88	0.88	1.00	0.85	0.85
MUC	0.97	0.95	0.91	0.97	0.94	0.90	0.97	0.96	0.92

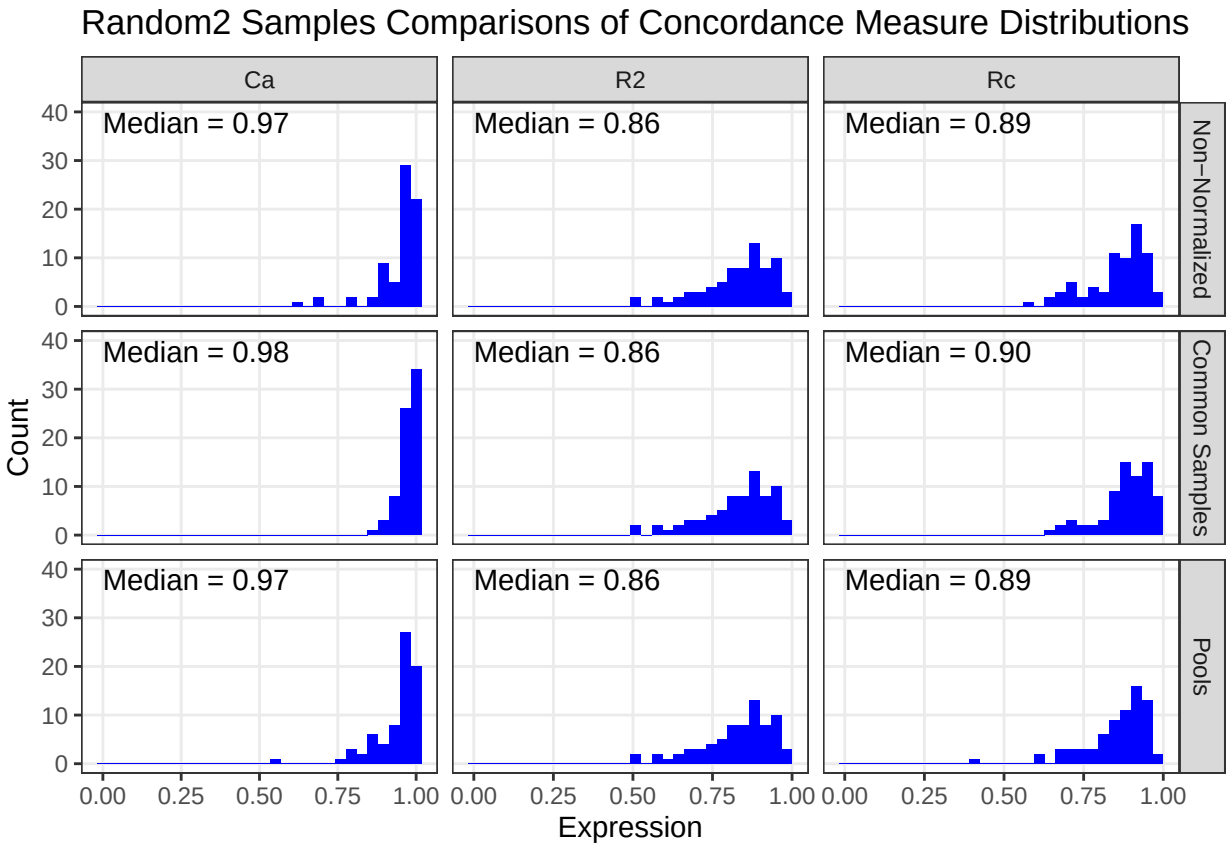


Figure 3.21: Random2 Samples Comparisons of Concordance Measure Distributions

Table 3.18: Random1 Samples Comparisons Statistics by Histotypes

hist	R2-Non	Ca-Non	Rc-Non	R2-Common	Ca-Common	Rc-Common	R2-Pools	Ca-Pools	Rc-Pools
CCOC	1.00	0.23	0.08	1.00	0.27	0.16	1.00	0.15	0.09
ENOC	1.00	0.63	0.61	1.00	0.61	0.57	1.00	0.61	0.61
HGSC	0.83	0.96	0.86	0.83	0.98	0.89	0.83	0.96	0.86
LGSC	0.98	0.92	0.90	0.98	0.95	0.93	0.98	0.92	0.90
MUC	0.68	0.77	0.55	0.68	0.86	0.61	0.68	0.78	0.51

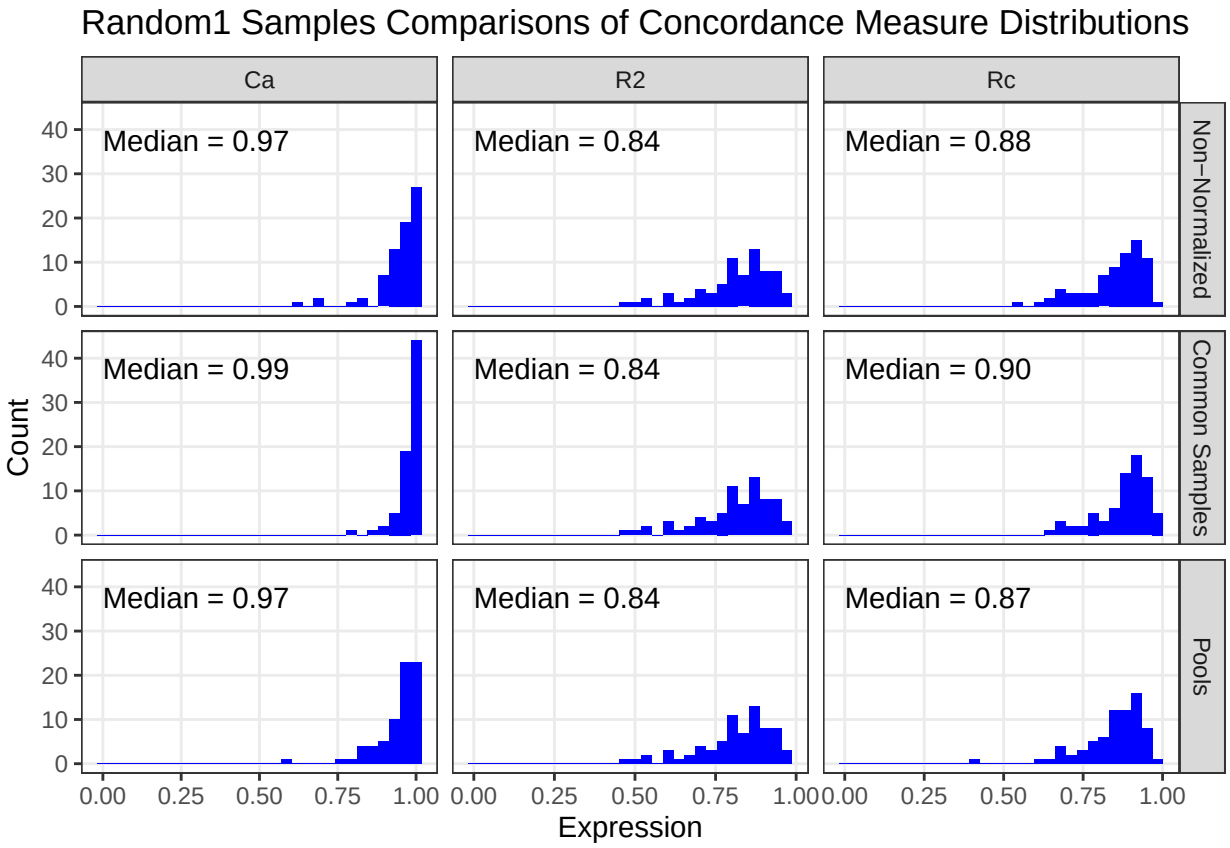


Figure 3.22: Random1 Samples Comparisons of Concordance Measure Distributions

3.3.4 CodeSet Chaining

3.3.4.1 CS1, CS2, CS3

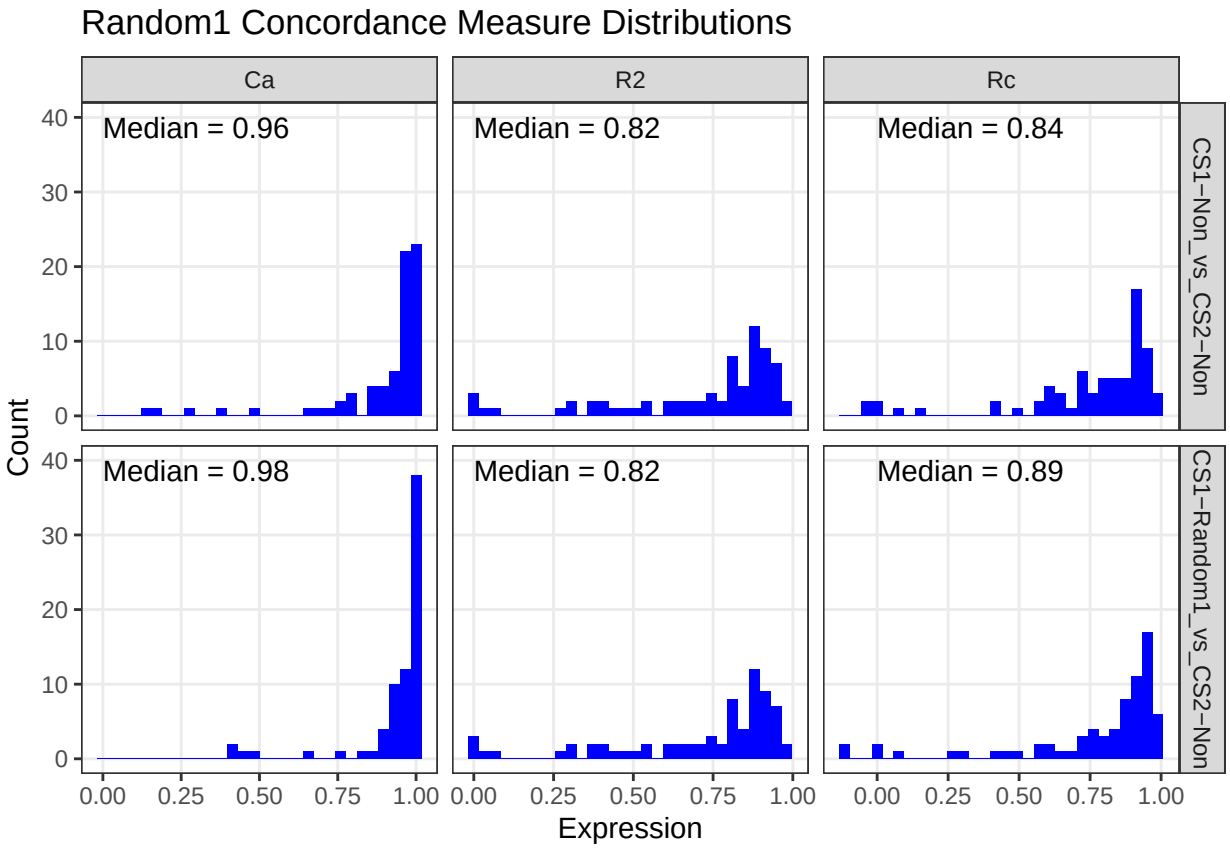


Figure 3.23: Random1 Concordance Measure Distributions

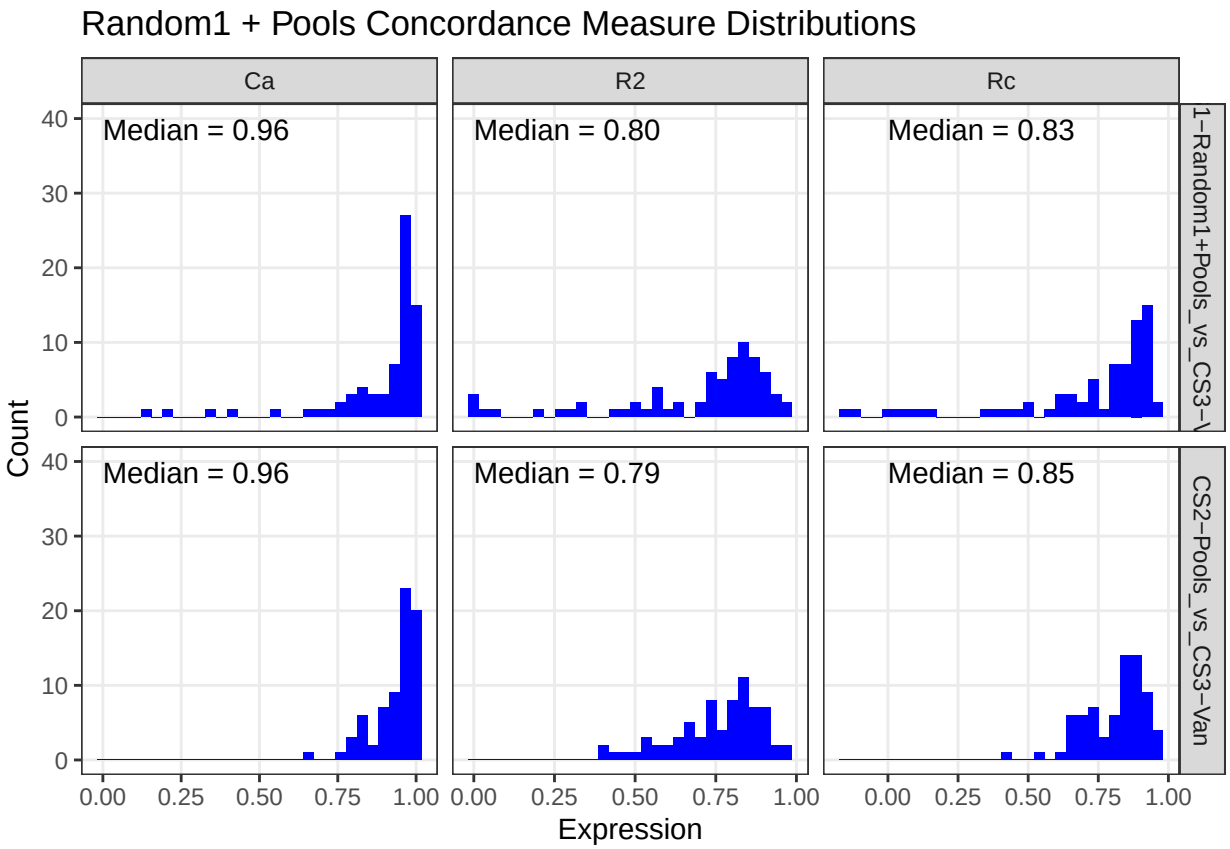


Figure 3.24: Random1 + Pools Concordance Measure Distributions

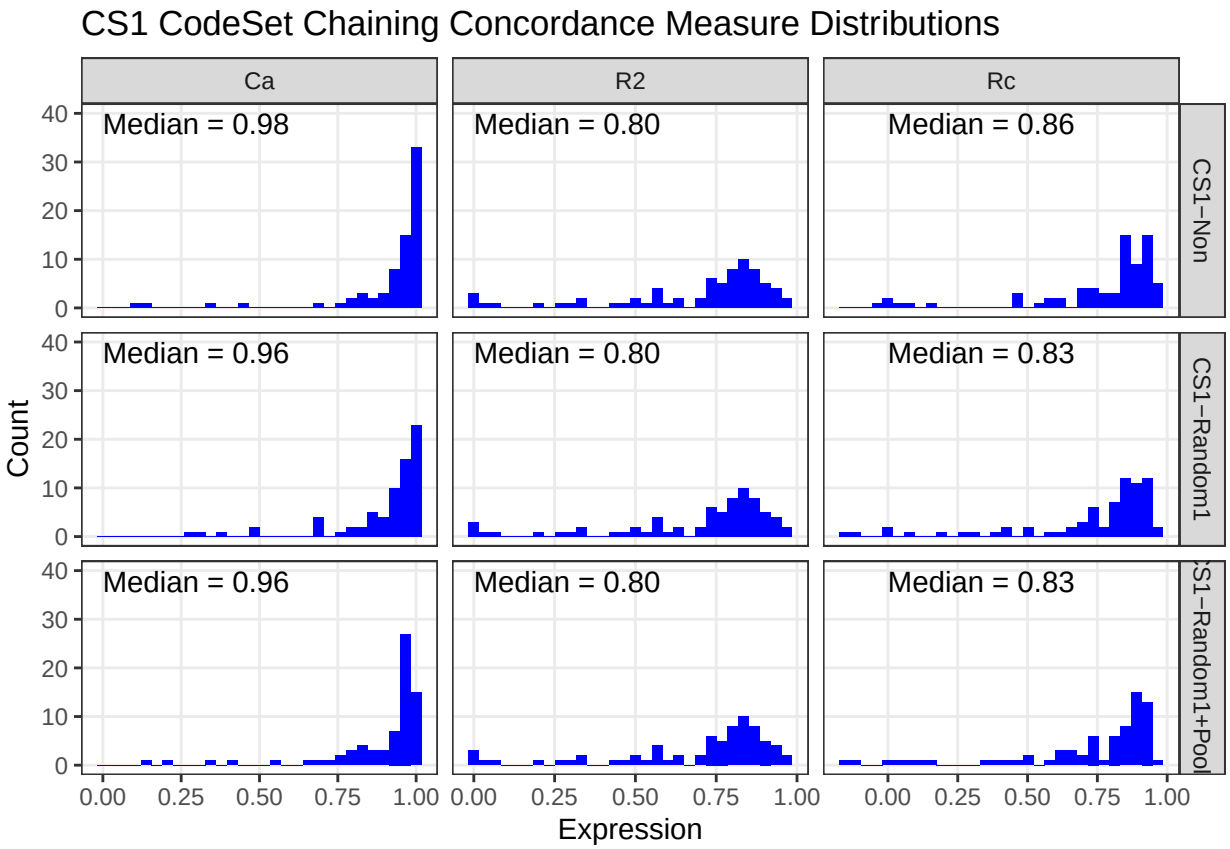


Figure 3.25: CS1 CodeSet Chaining Concordance Measure Distributions

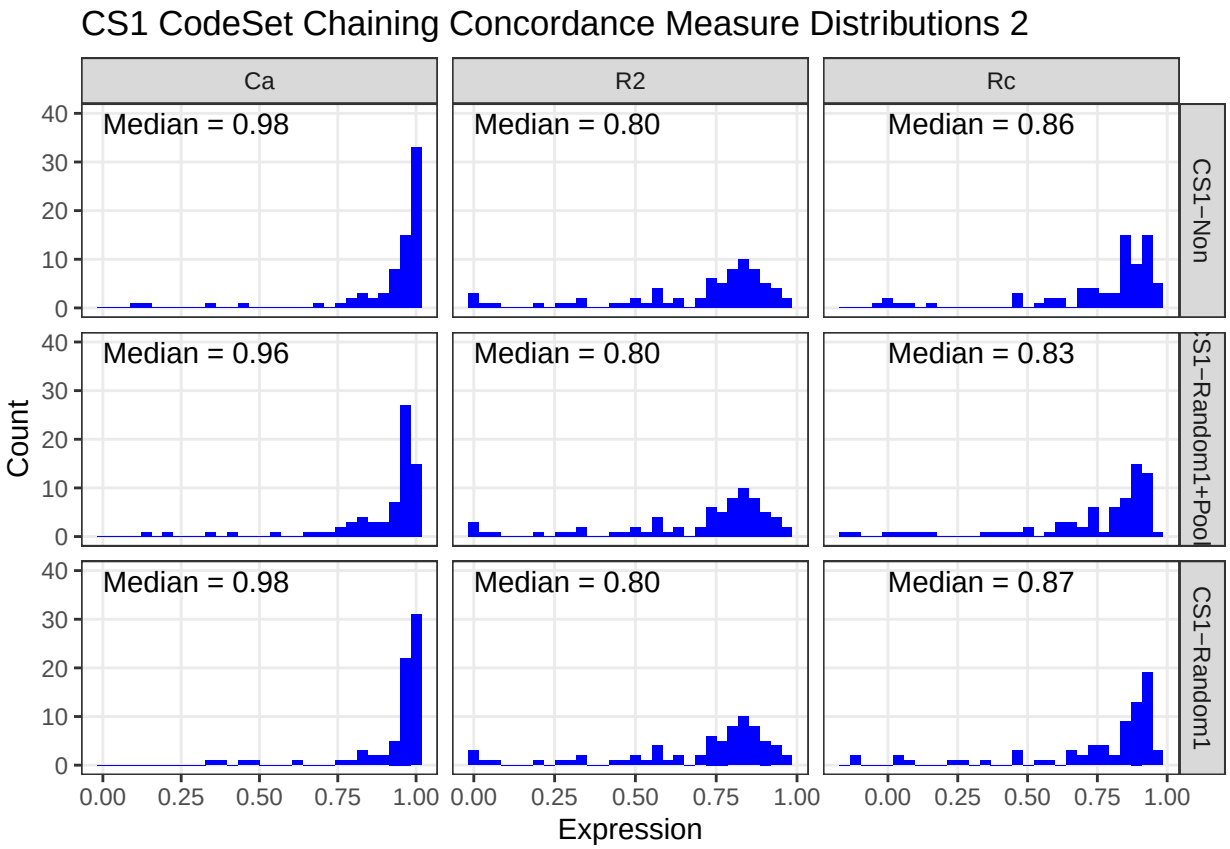


Figure 3.26: CS1 CodeSet Chaining Concordance Measure Distributions 2

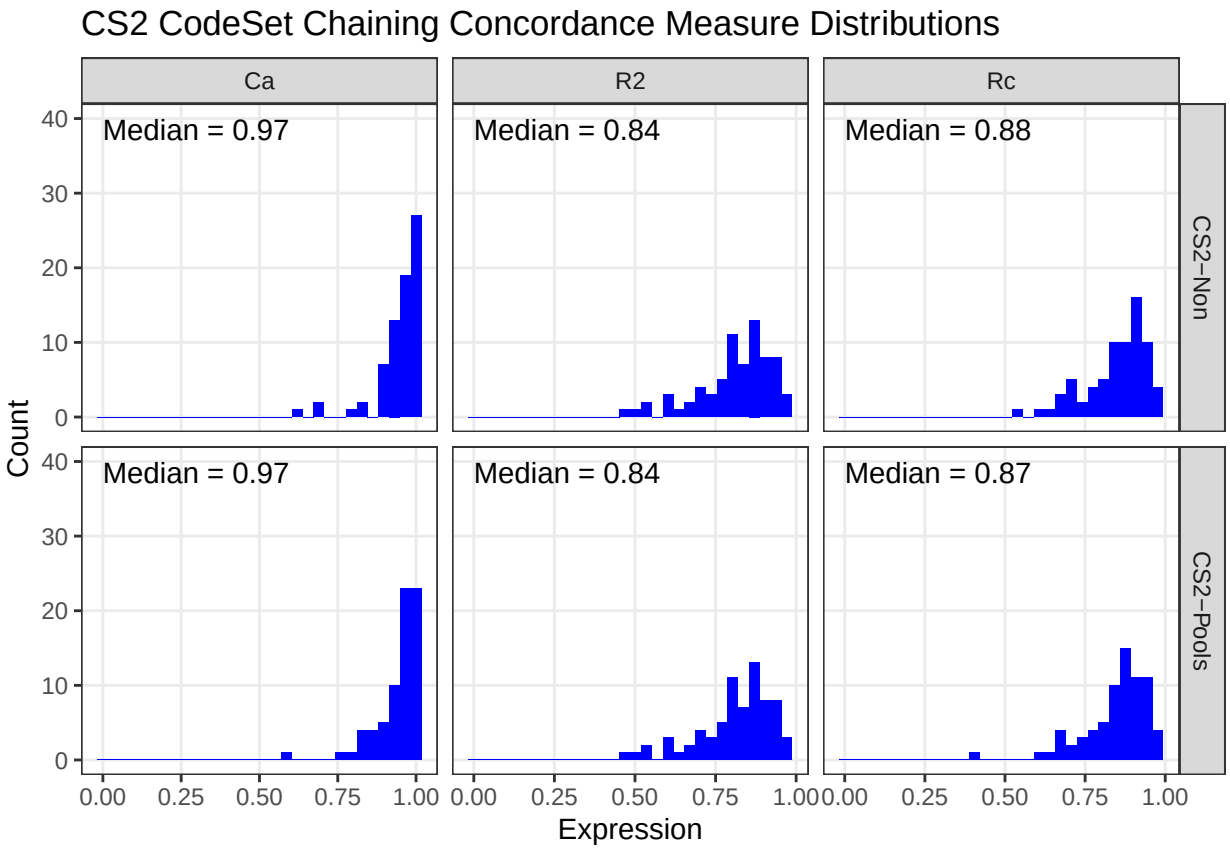


Figure 3.27: CS2 CodeSet Chaining Concordance Measure Distributions

3.3.4.2 CS3, CS4, CS5 using Set B/A

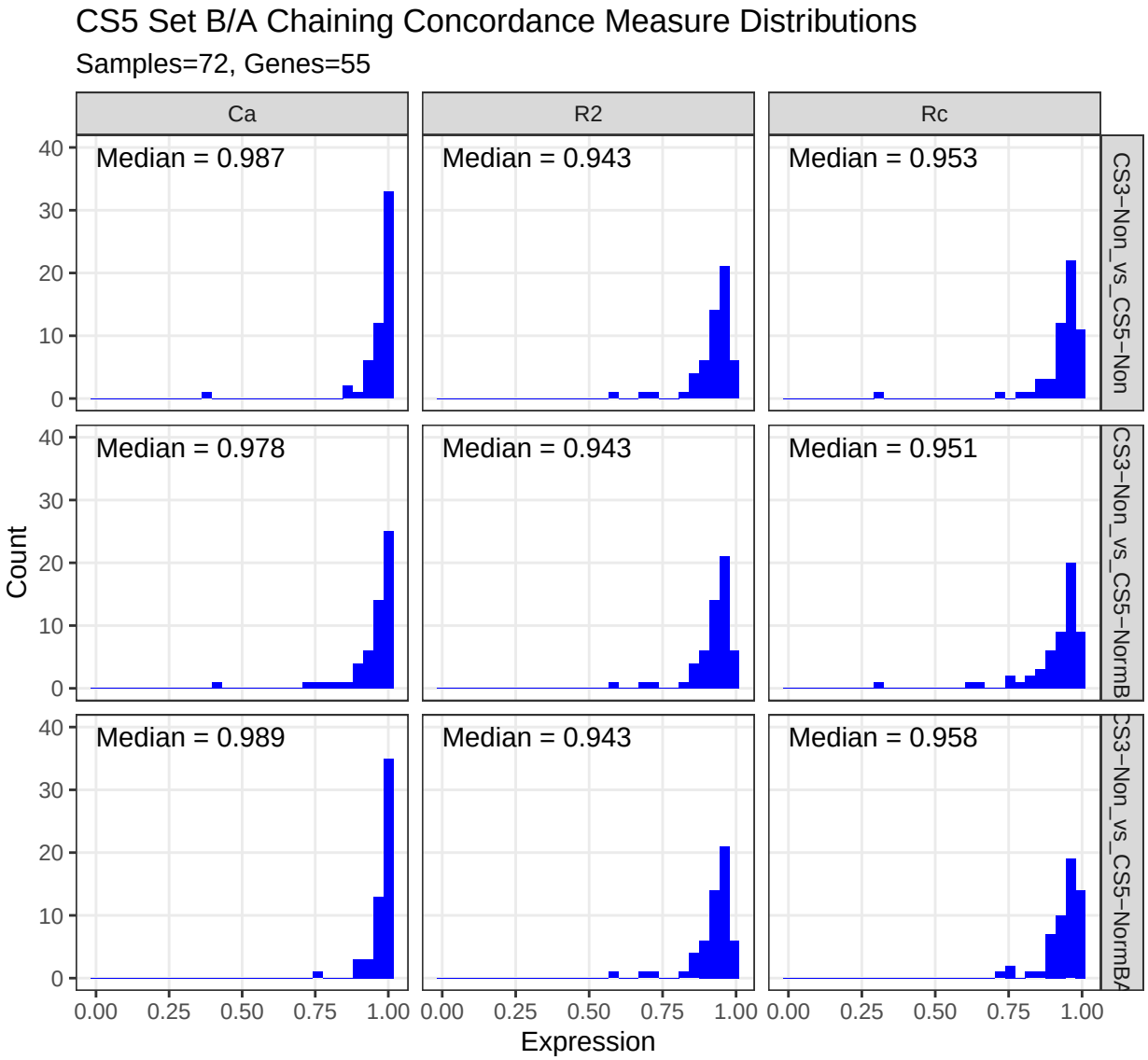


Figure 3.28: CS5 Set B/A Chaining Concordance Measure Distributions

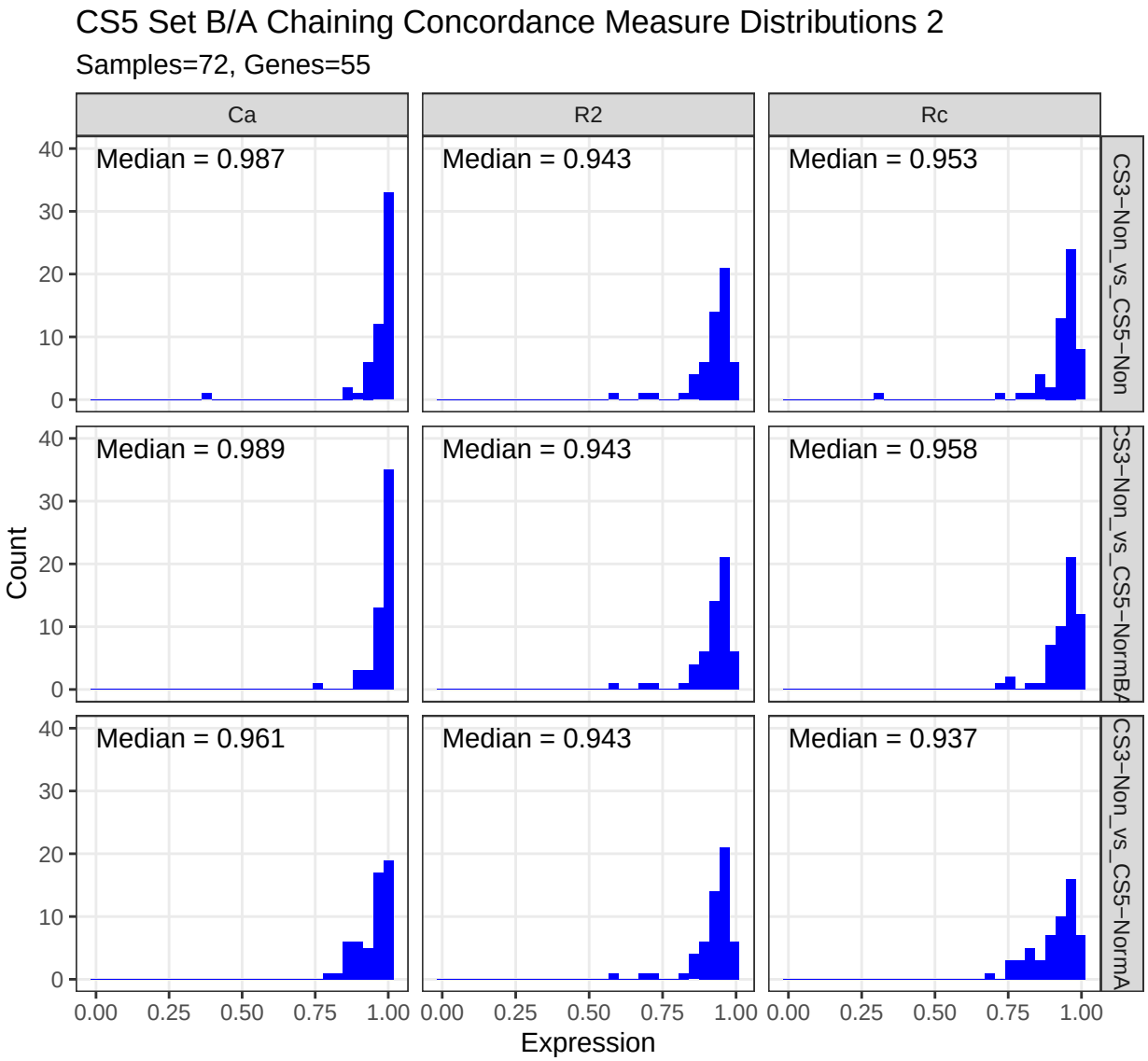


Figure 3.29: CS5 Set B/A Chaining Concordance Measure Distributions 2

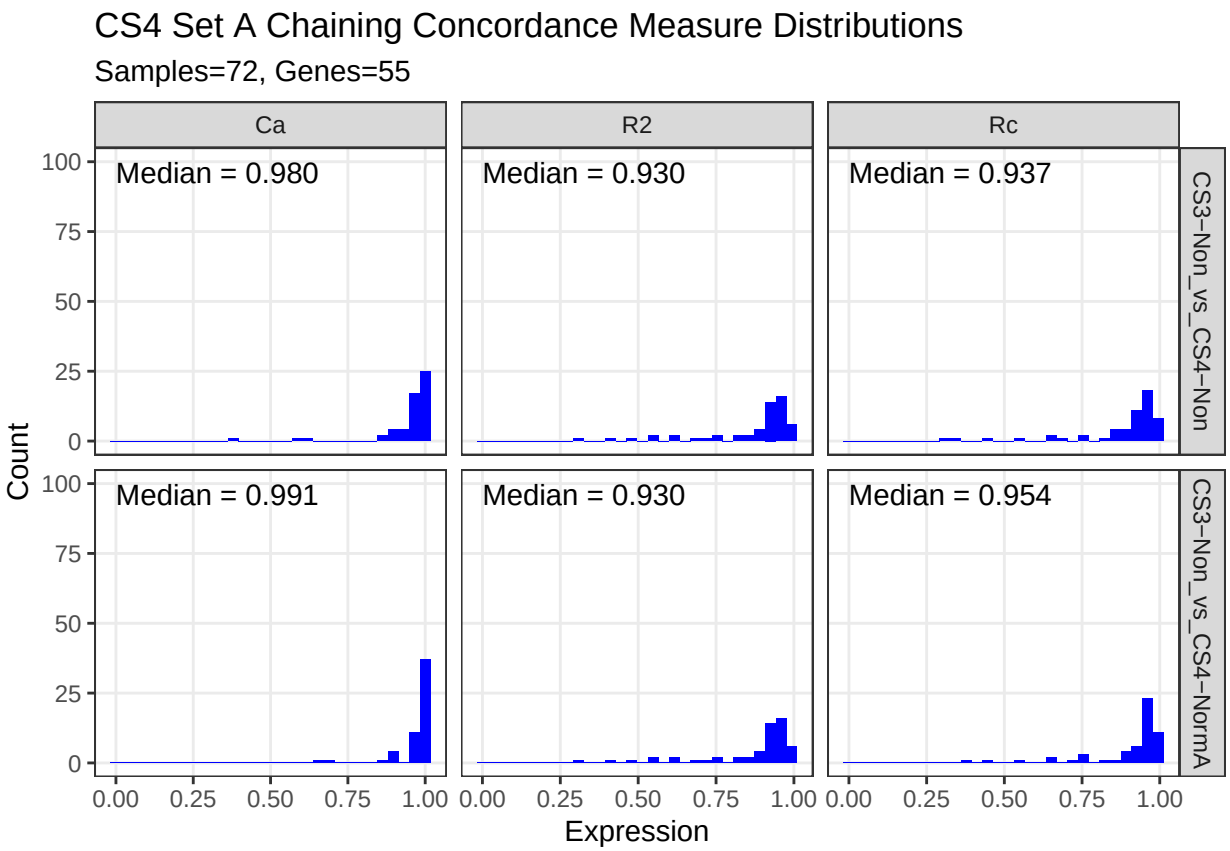


Figure 3.30: CS4 Set A Chaining Concordance Measure Distributions

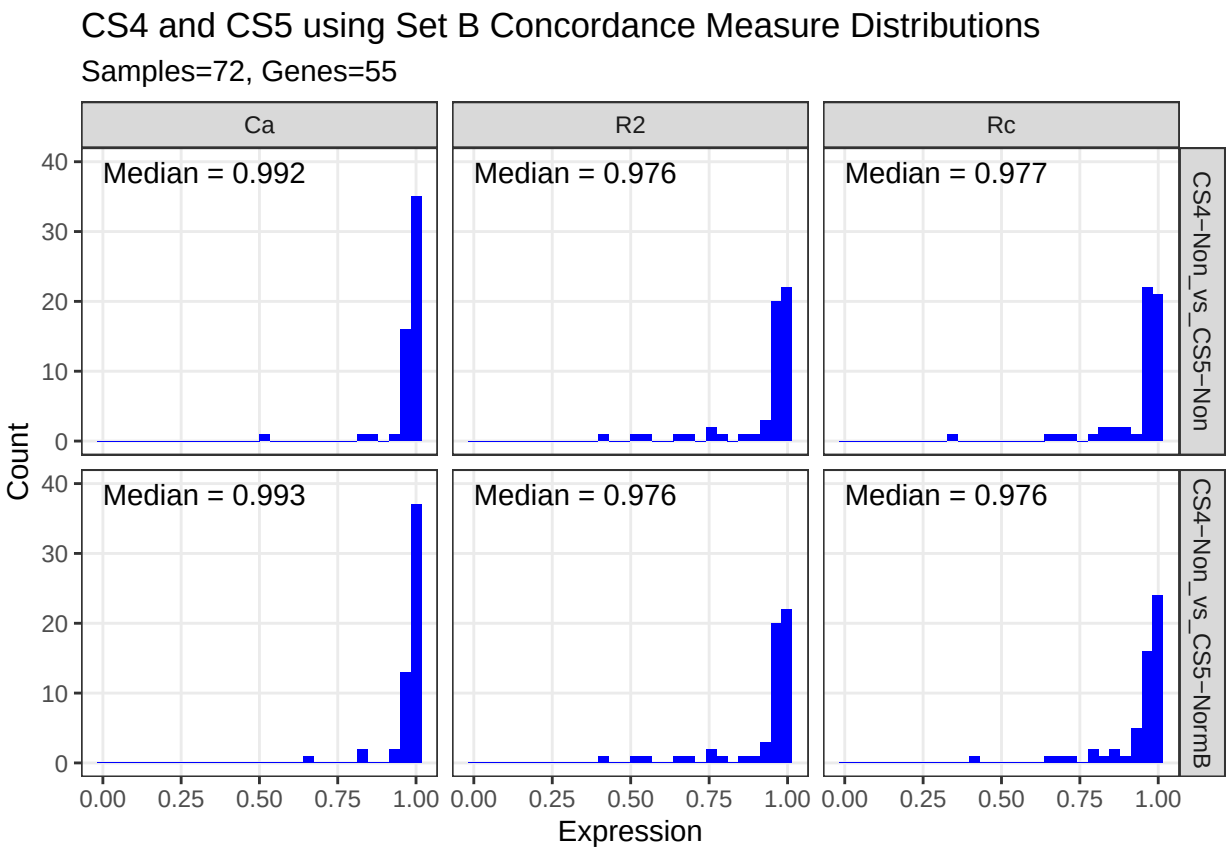


Figure 3.31: CS4 and CS5 using Set B Concordance Measure Distributions

3.3.4.3 CS3, CS4, CS5 using Set C/A

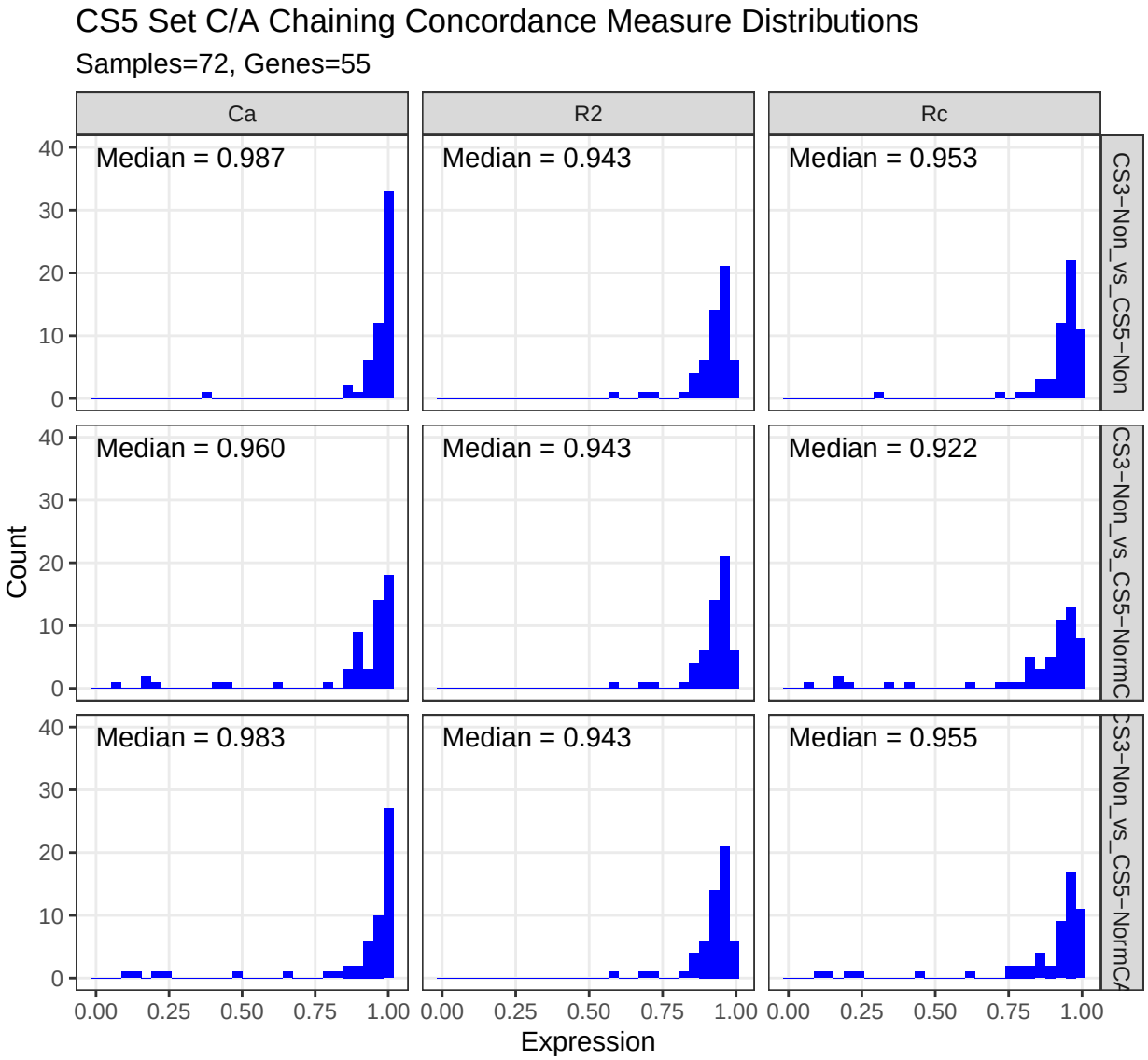


Figure 3.32: CS5 Set C/A Chaining Concordance Measure Distributions

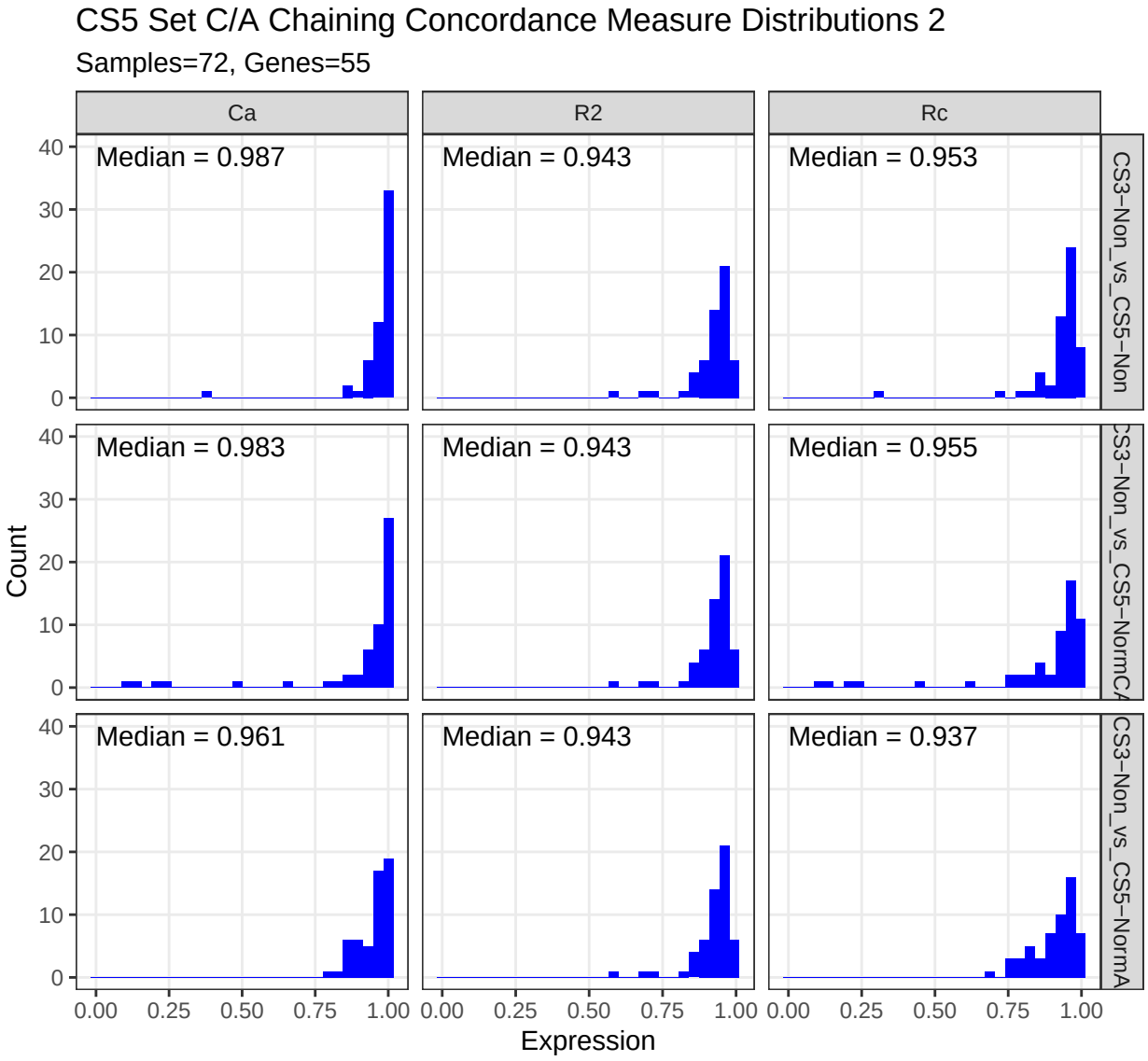


Figure 3.33: CS5 Set C/A Chaining Concordance Measure Distributions 2

Table 3.19: All Common Samples Histotype Distribution

revHist	CS1	CS2	CS3
CCOC	3	4	3
ENOC	4	4	3
HGSC	59	62	75
LGSC	7	5	4
MUC	7	5	5

Table 3.20: Distinct Common Samples Histotype Distribution

revHist	CS1	CS2	CS3
CCOC	3	3	3
ENOC	3	3	3
HGSC	57	57	57
LGSC	4	4	4
MUC	5	5	5

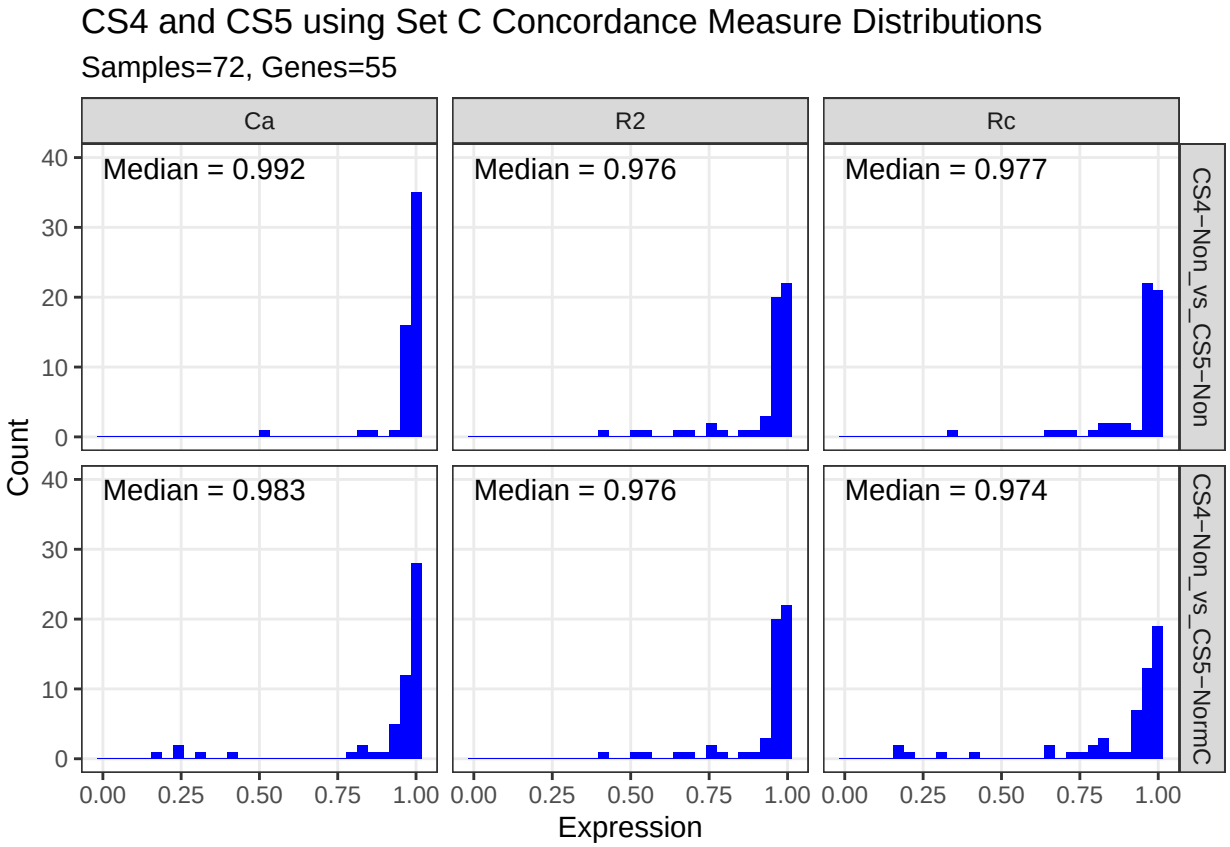


Figure 3.34: CS4 and CS5 using Set C Concordance Measure Distributions

3.4 Common Sample Distributions

Table 3.21: Distinct Common CS2 and CS3 Samples Histotype Distribution

revHist	CS2	CS3
CCOC	3	3
ENOC	3	3
HGSC	71	71
LGSC	4	4
MUC	5	5

Table 3.22: Common Samples Across Sites Histotype Distribution

revHist	AOC	USC	Vancouver
CCOC	3	3	3
ENOC	3	3	3
HGSC	13	13	27
LGSC	2	2	2
MUC	3	3	3

Table 3.23: Distinct Common Samples Across Sites Histotype Distribution

revHist	AOC	USC	Vancouver
CCOC	3	3	3
ENOC	3	3	3
HGSC	13	13	13
LGSC	2	2	2
MUC	3	3	3

Table 3.24: CS3/CS4/CS5 Common Samples Histotype Distribution

revHist	CS3	CS4	CS5
HGSC	46	46	46
NA	26	26	26

Table 3.25: CS3/CS4/CS5 Pools Distribution

Pool	CS3	CS4	CS5
Pool1	12	5	4
Pool2	5	5	4
Pool3	5	5	4
Pool4	NA	2	1
Pool5	NA	2	1
Pool6	NA	2	0
Pool7	NA	2	1
Pool8	NA	2	1
Pool9	NA	2	1
Pool10	NA	2	1
Pool11	NA	2	1

Table 3.26: Full Training Set Histotype Distribution

revHist	n	freq
HGSC	1227	79%
CCOC	106	7%
ENOC	91	6%
MUC	84	5%
LGSC	39	3%

Table 3.27: Full Training Set Histotype Distribution by CodeSet

Variable	Levels	CS1	CS2	CS3	Total
Histotype	HGSC	122 (49%)	629 (80%)	476 (94%)	1227 (79%)
	CCOC	44 (18%)	54 (7%)	8 (2%)	106 (7%)
	ENOC	55 (22%)	28 (4%)	8 (2%)	91 (6%)
	MUC	16 (6%)	59 (7%)	9 (2%)	84 (5%)
	LGSC	14 (6%)	19 (2%)	6 (1%)	39 (3%)
Total	N (%)	251 (16%)	789 (51%)	507 (33%)	1547 (100%)

3.5 Histotype Distribution in Classifier Datasets

Table 3.28: CS1 All Training Set Histotype Distribution

revHist	n	freq
HGSC	125	47%
ENOC	58	22%
CCOC	47	18%
LGSC	19	7%
MUC	19	7%

Table 3.29: CS2 All Training Set Histotype Distribution

revHist	n	freq
HGSC	654	79%
MUC	61	7%
CCOC	60	7%
ENOC	32	4%
LGSC	20	2%

Table 3.30: Confirmation Set Histotype Distribution

revHist	n	freq
HGSC	423	66%
ENOC	106	16%
CCOC	75	12%
MUC	27	4%
LGSC	13	2%

Table 3.31: Validation Set Histotype Distribution

revHist	n	freq
HGSC	781	74%
ENOC	140	13%
CCOC	86	8%
MUC	34	3%
LGSC	20	2%

4. Results

We show internal validation summaries for the combined classifier training set, as well as the CS1 and CS2 sets with duplicates included. The F1-scores, kappa, and G-mean are the measures of interest. Algorithms are sorted by descending value based on the overallaccuracy of the training set. The point ranges show the median, 5th and 95th percentiles, coloured by subsampling methods.

4.1 Training Set

4.1.1 Accuracy

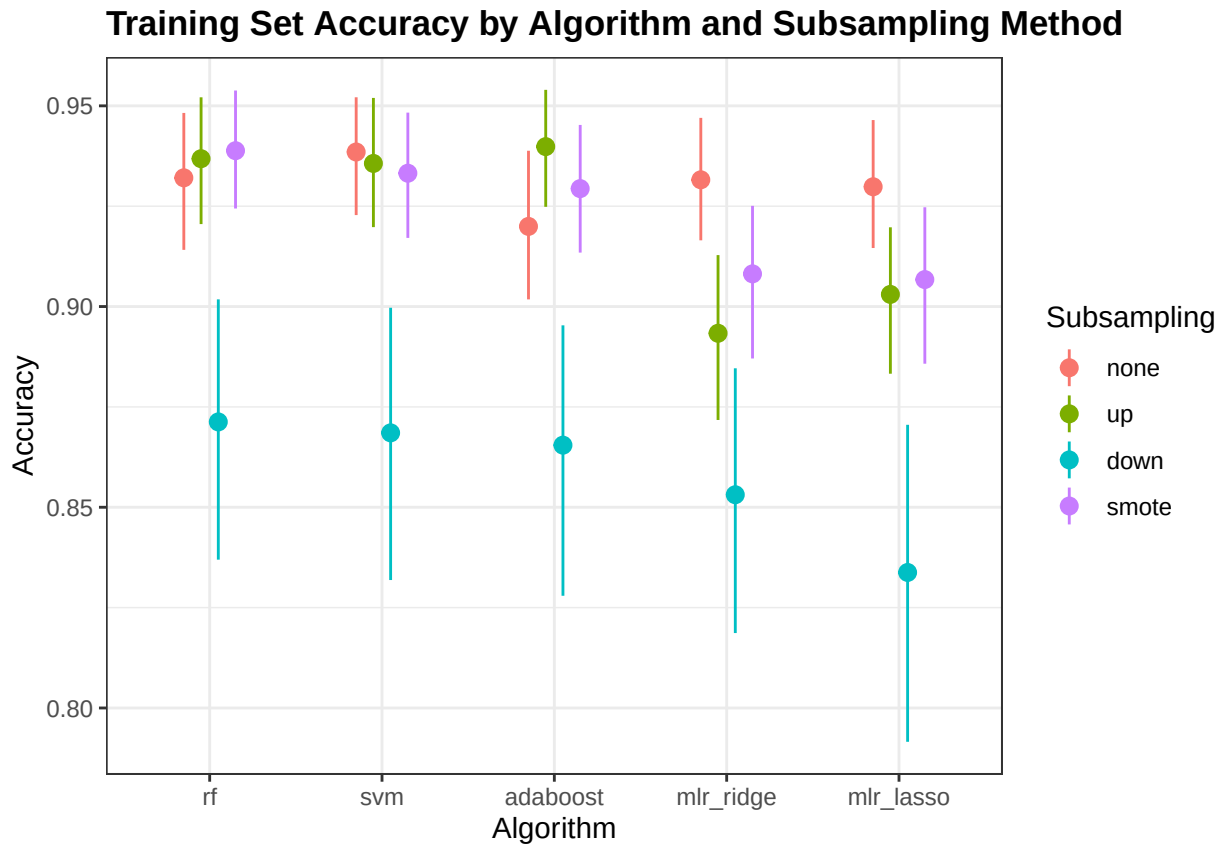


Figure 4.1: Training Set Accuracy

4.1.2 F1-Score

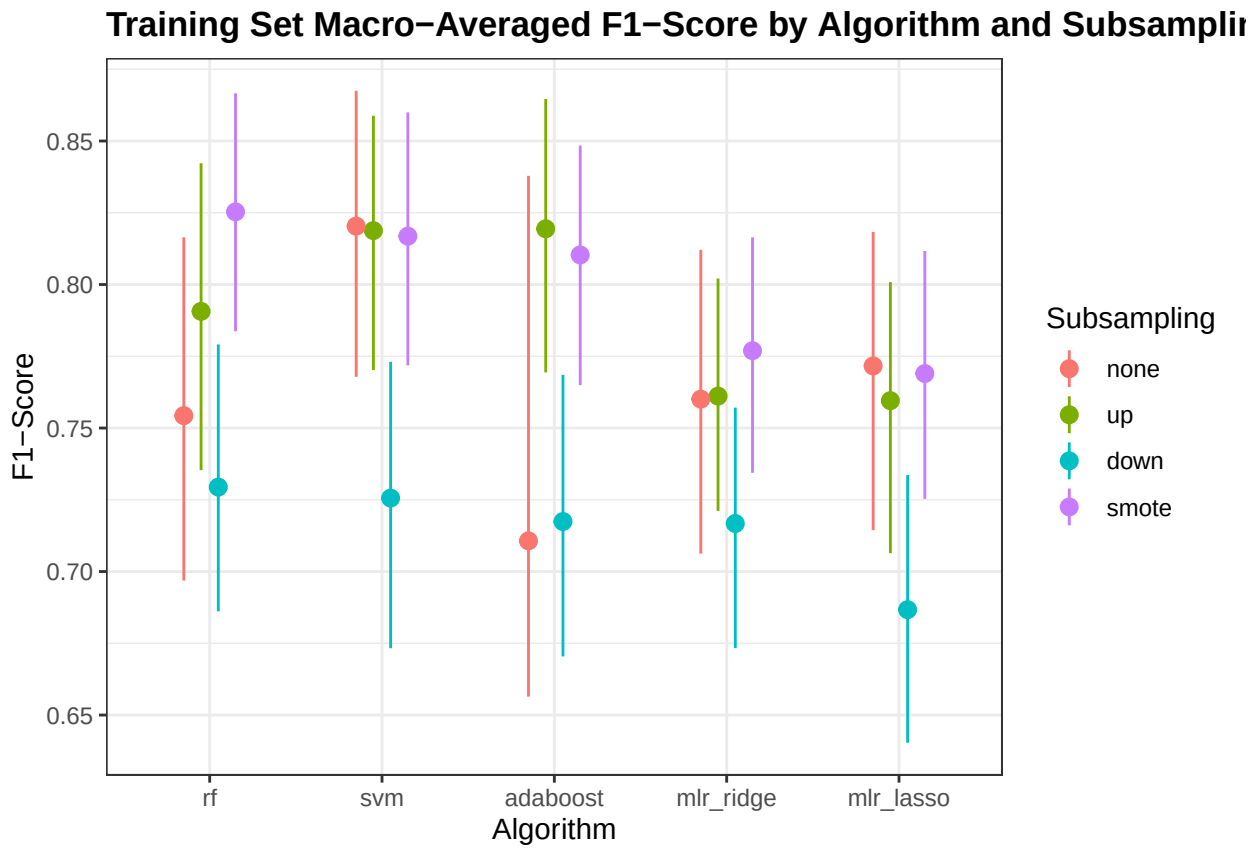


Figure 4.2: Training Set F1-Score

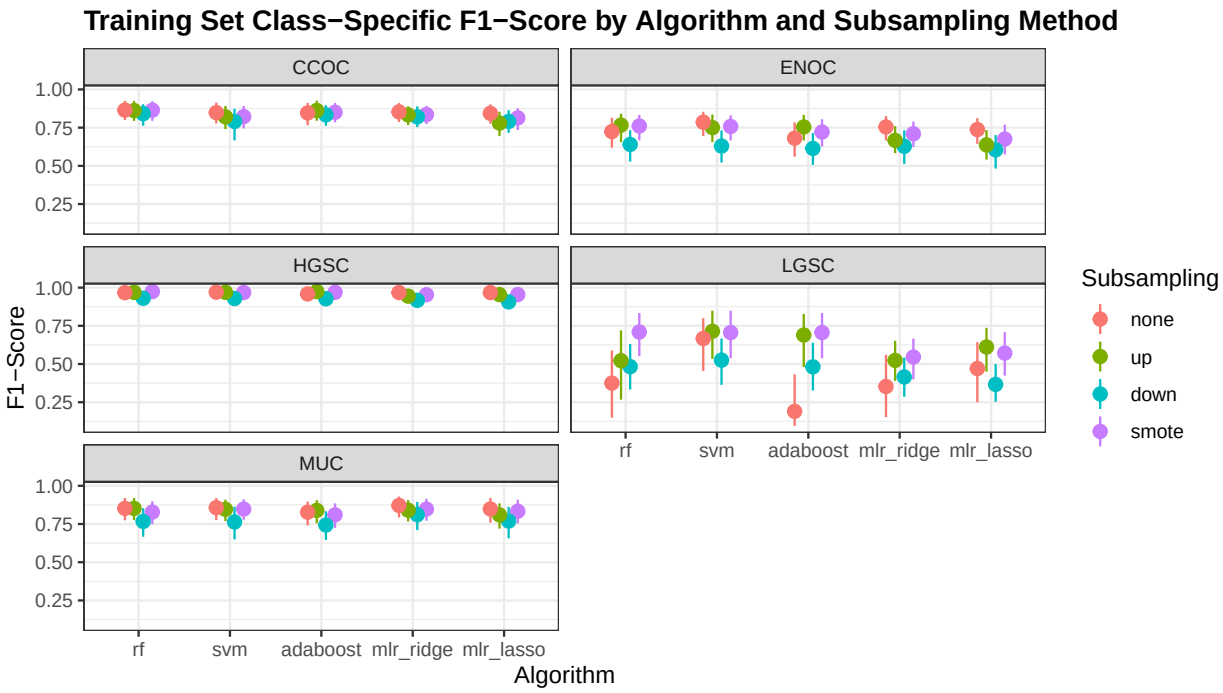


Figure 4.3: Training Set Class-Specific F1-Score

Table 4.1: Training Set Kappa by Algorithm and Subsampling Method

sampling	rf	svm	adaboost	mlr_ridge	mlr_lasso
none	0.793	0.819	0.745	0.798	0.798
up	0.809	0.808	0.828	0.739	0.745
down	0.698	0.691	0.683	0.667	0.631
smote	**0.831**	0.811	0.81	0.766	0.758

4.1.3 Kappa

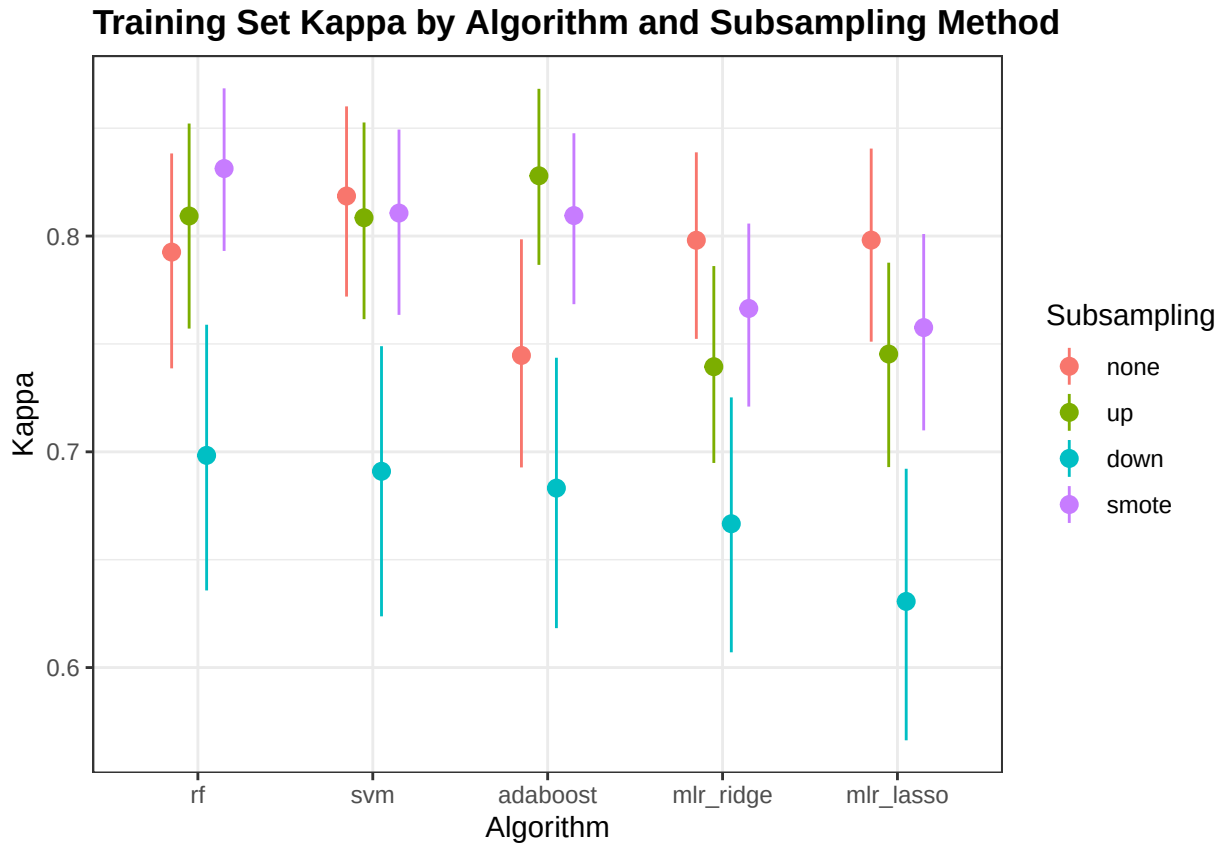


Figure 4.4: Training Set Kappa

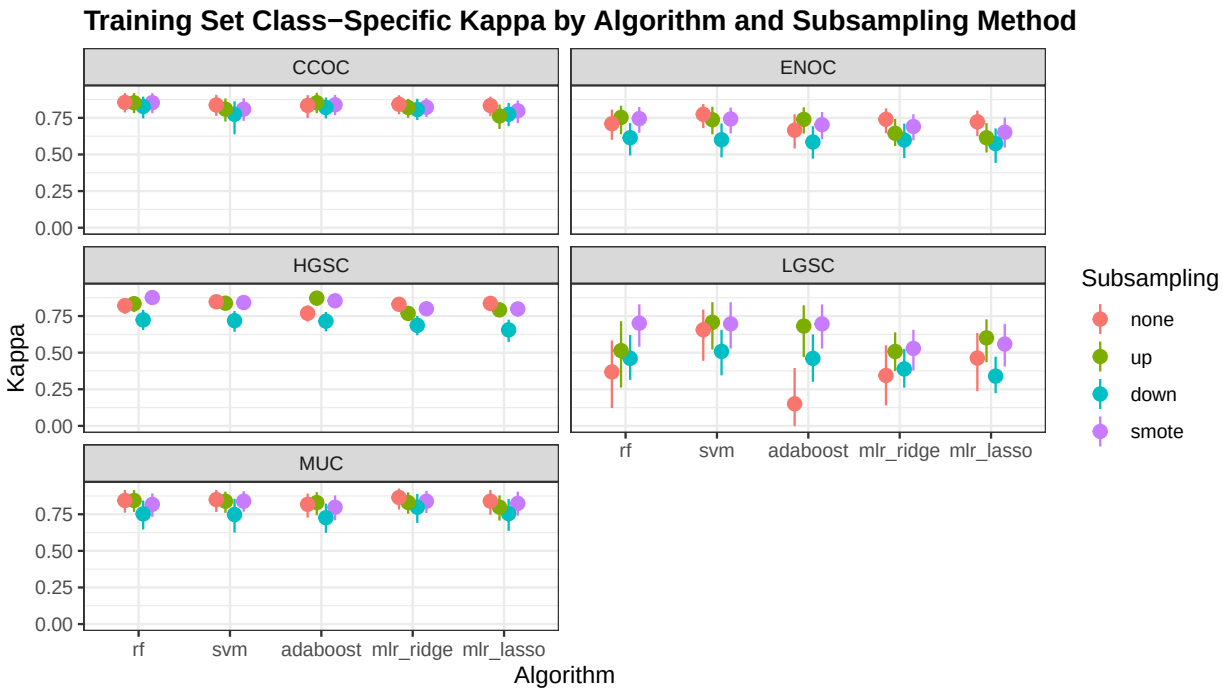


Figure 4.5: Training Set Class-Specific Kappa

4.1.4 G-mean

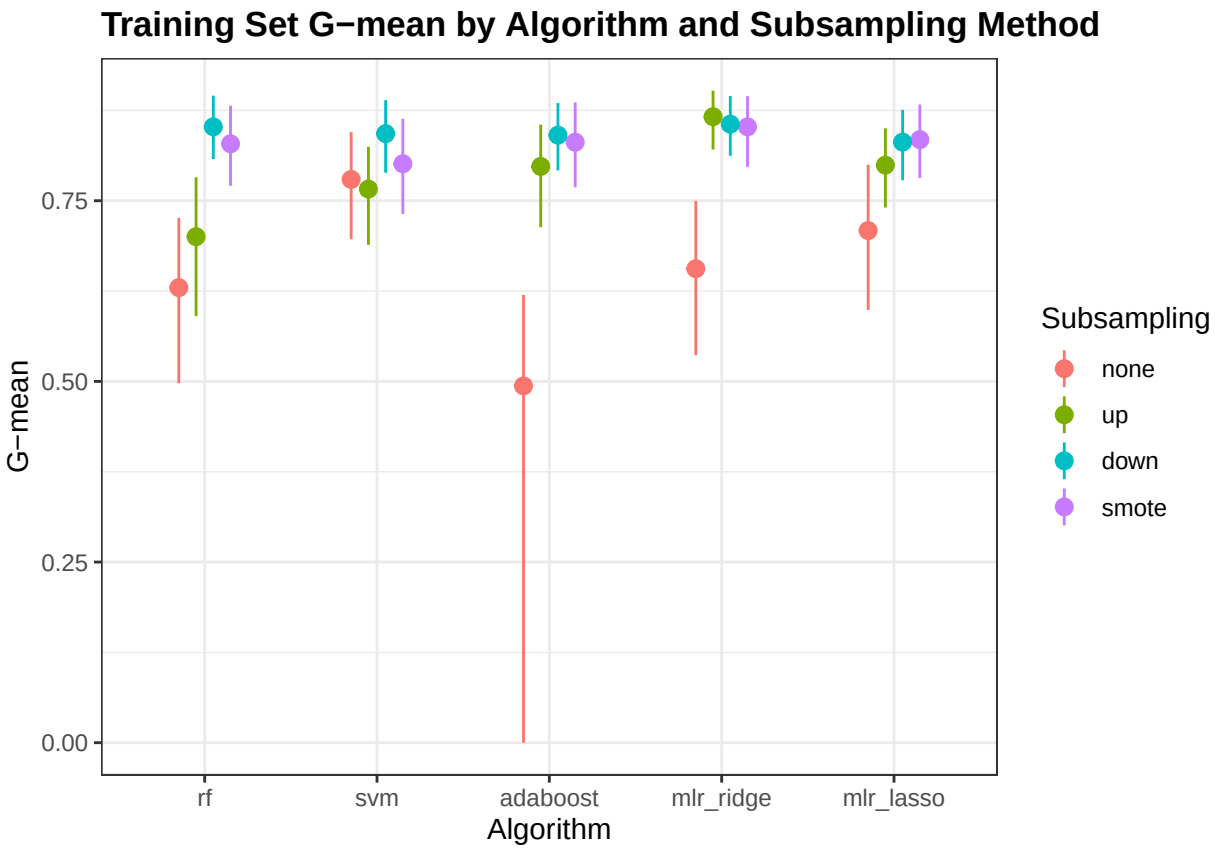


Figure 4.6: Training Set G-mean

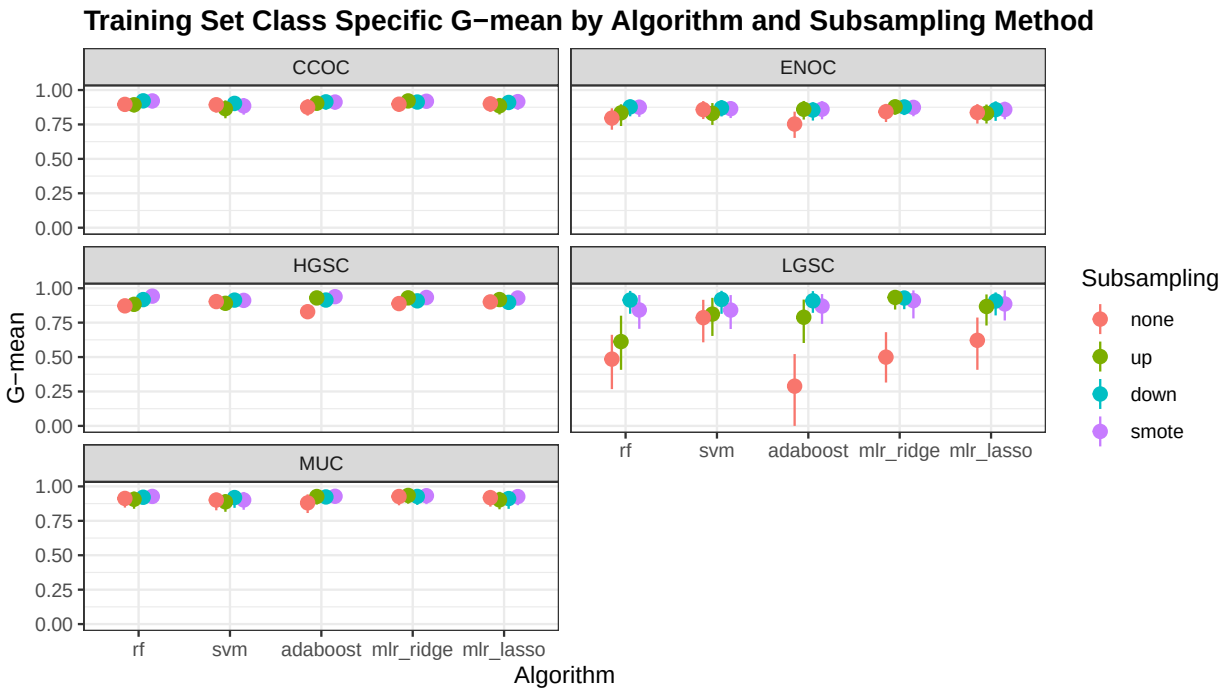


Figure 4.7: Training Set Class-Specific G-mean

4.2 CS1 Set

4.2.1 Accuracy

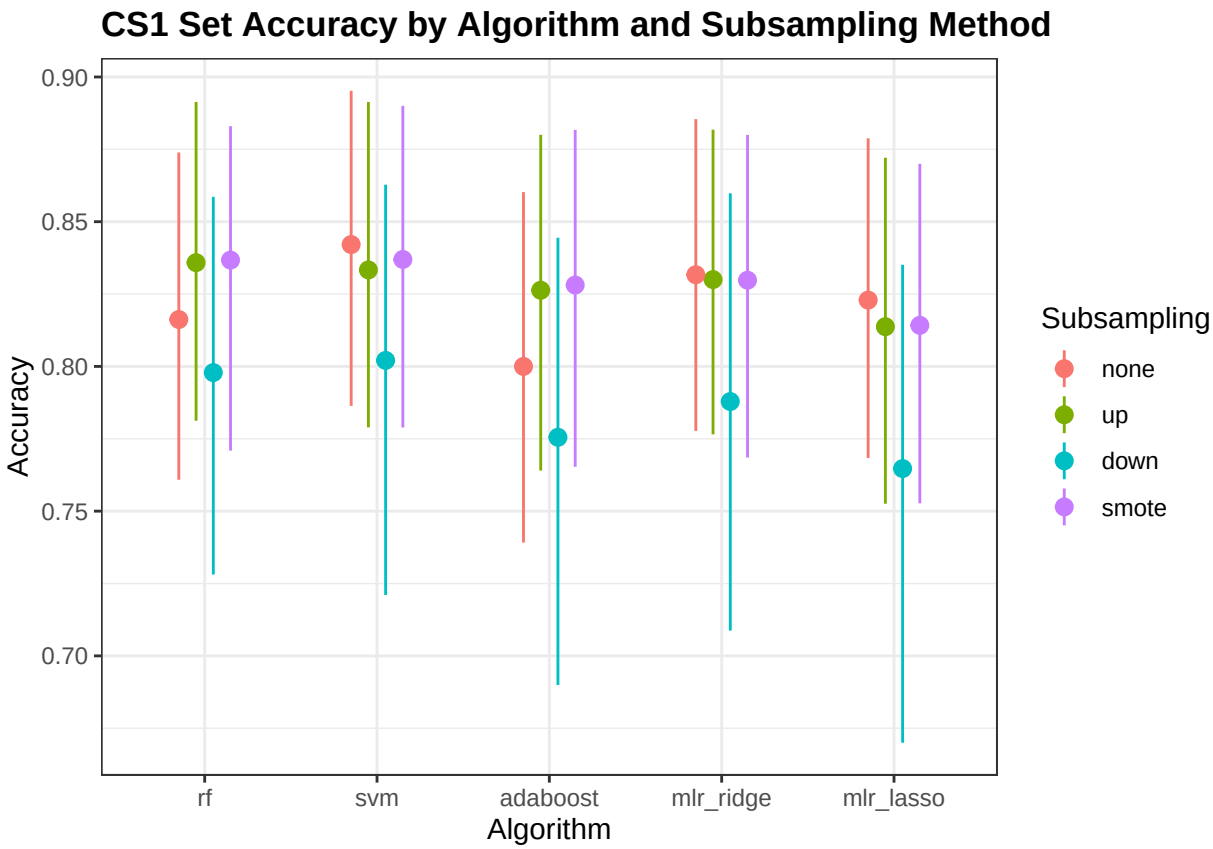


Figure 4.8: CS1 Set Accuracy

4.2.2 F1-Score

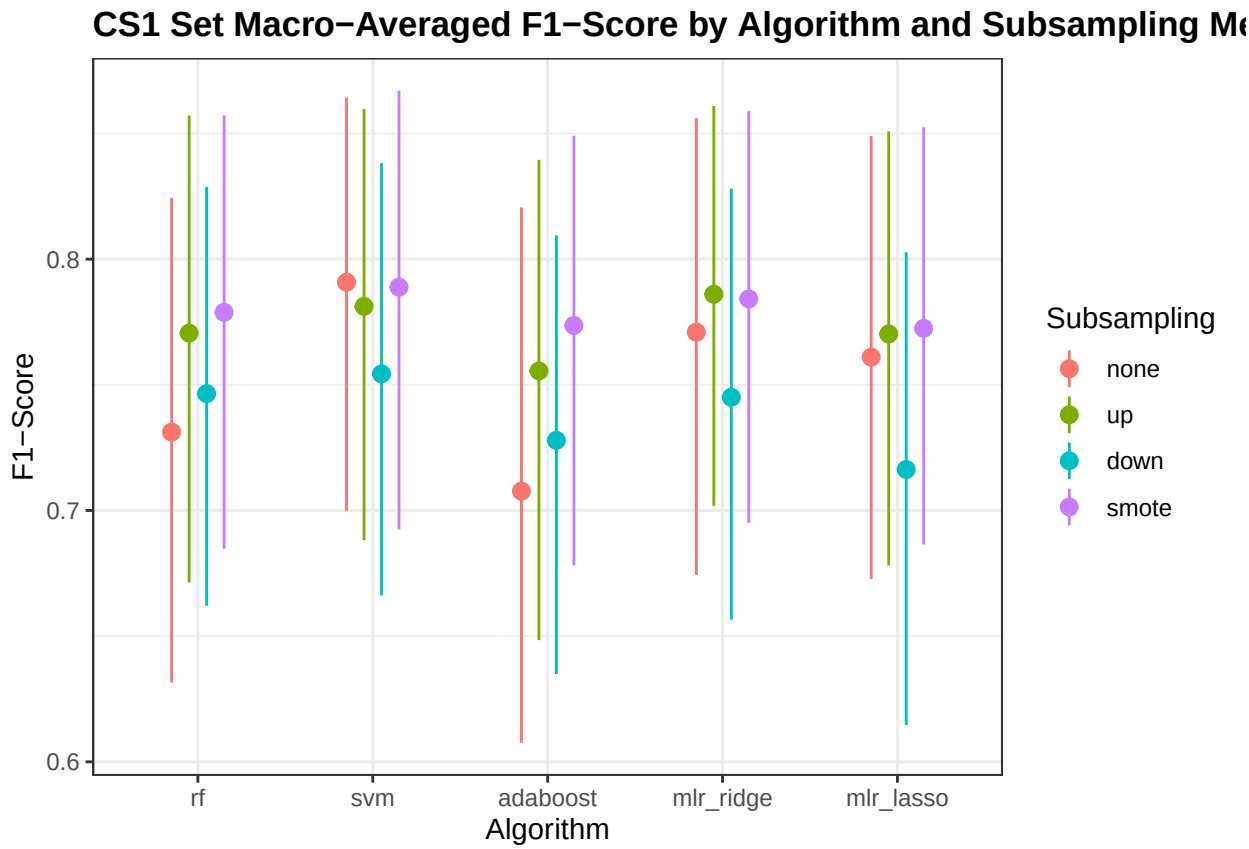


Figure 4.9: CS1 Set F1-Score

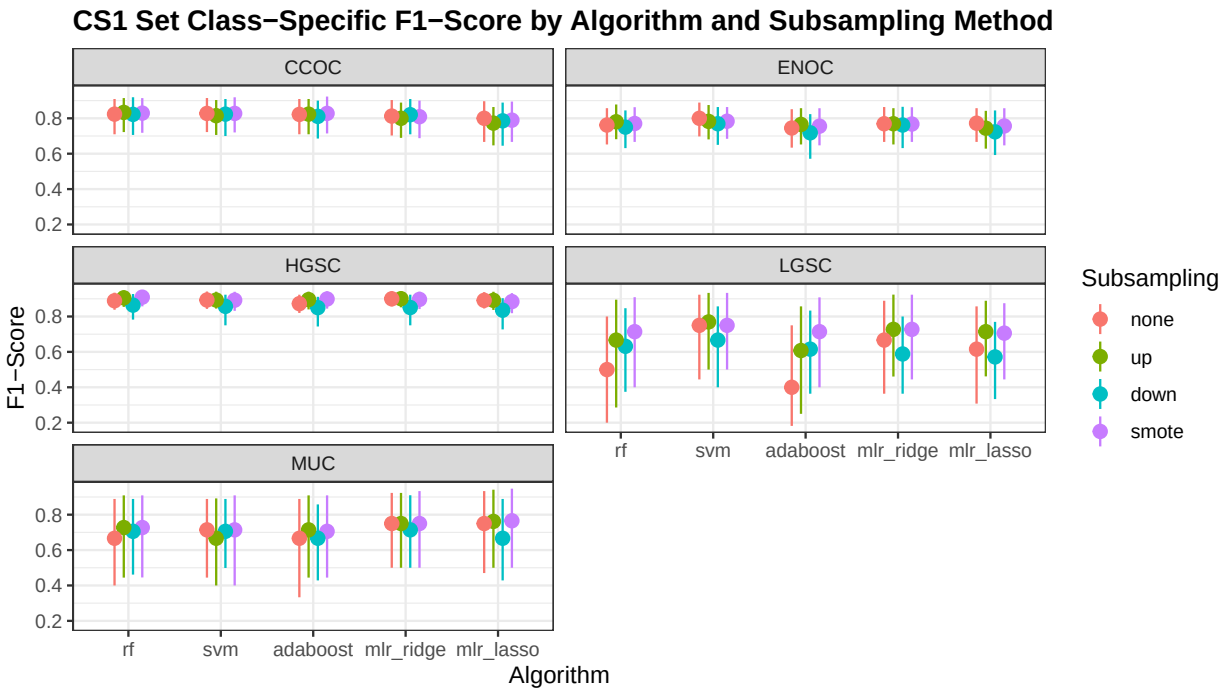


Figure 4.10: CS1 Set Class-Specific F1-Score

Table 4.2: CS1 Set Kappa by Algorithm and Subsampling Method

sampling	rf	svm	adaboost	mlr_ridge	mlr_lasso
none	0.724	**0.766**	0.697	0.752	0.74
up	0.756	0.752	0.742	0.757	0.732
down	0.716	0.72	0.687	0.706	0.674
smote	0.763	0.759	0.751	0.755	0.736

4.2.3 Kappa

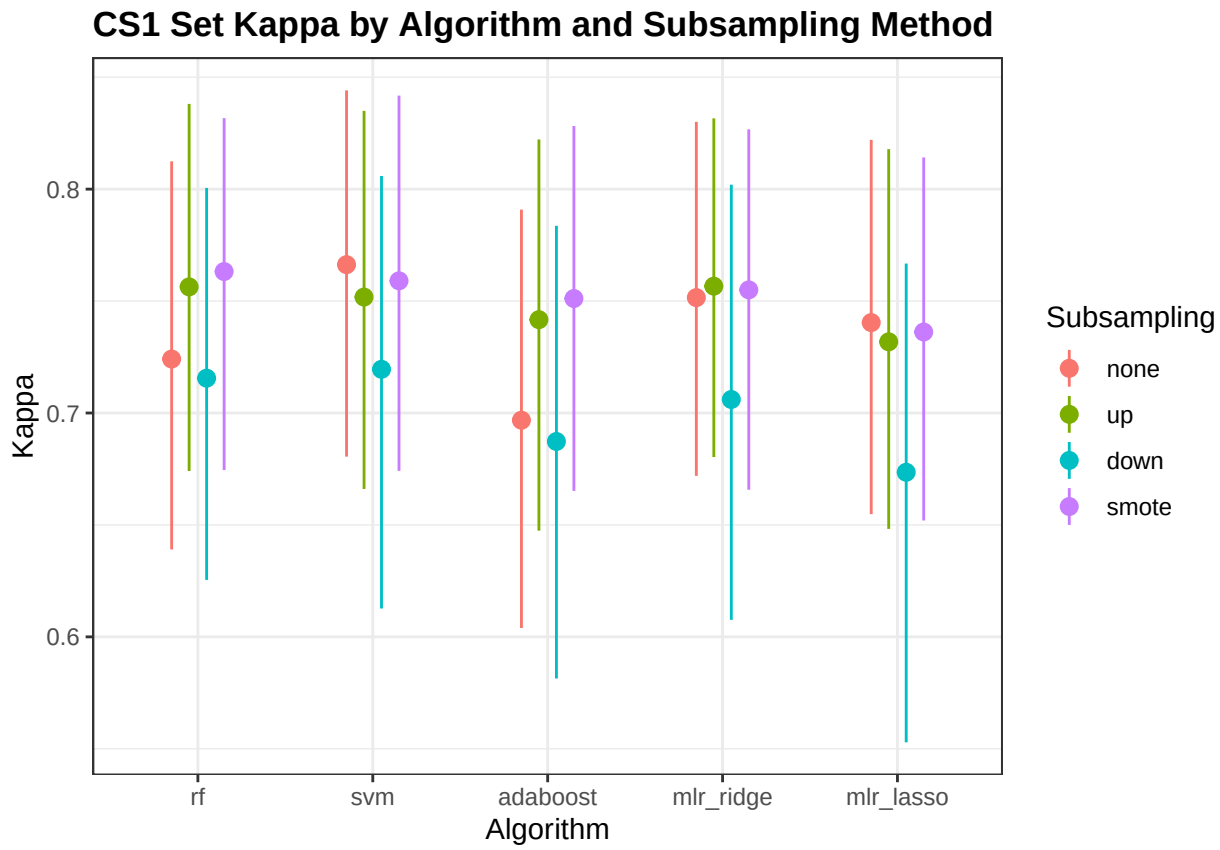


Figure 4.11: CS1 Set Kappa

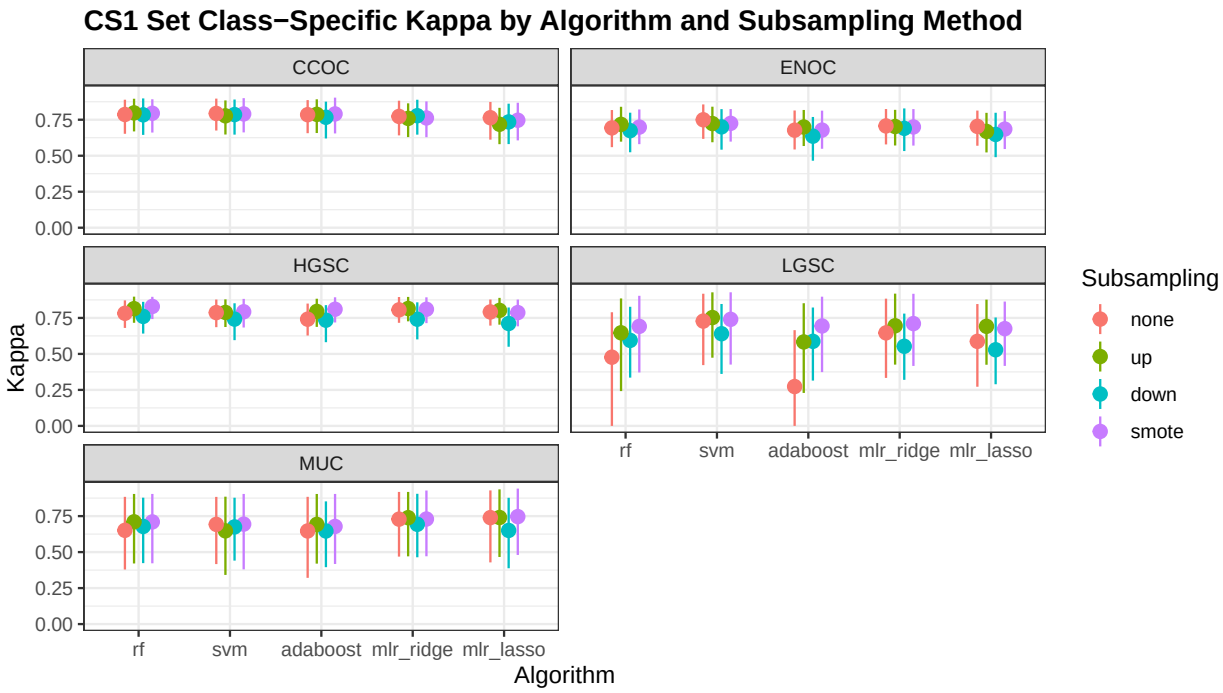


Figure 4.12: CS1 Set Class-Specific Kappa

4.2.4 G-mean

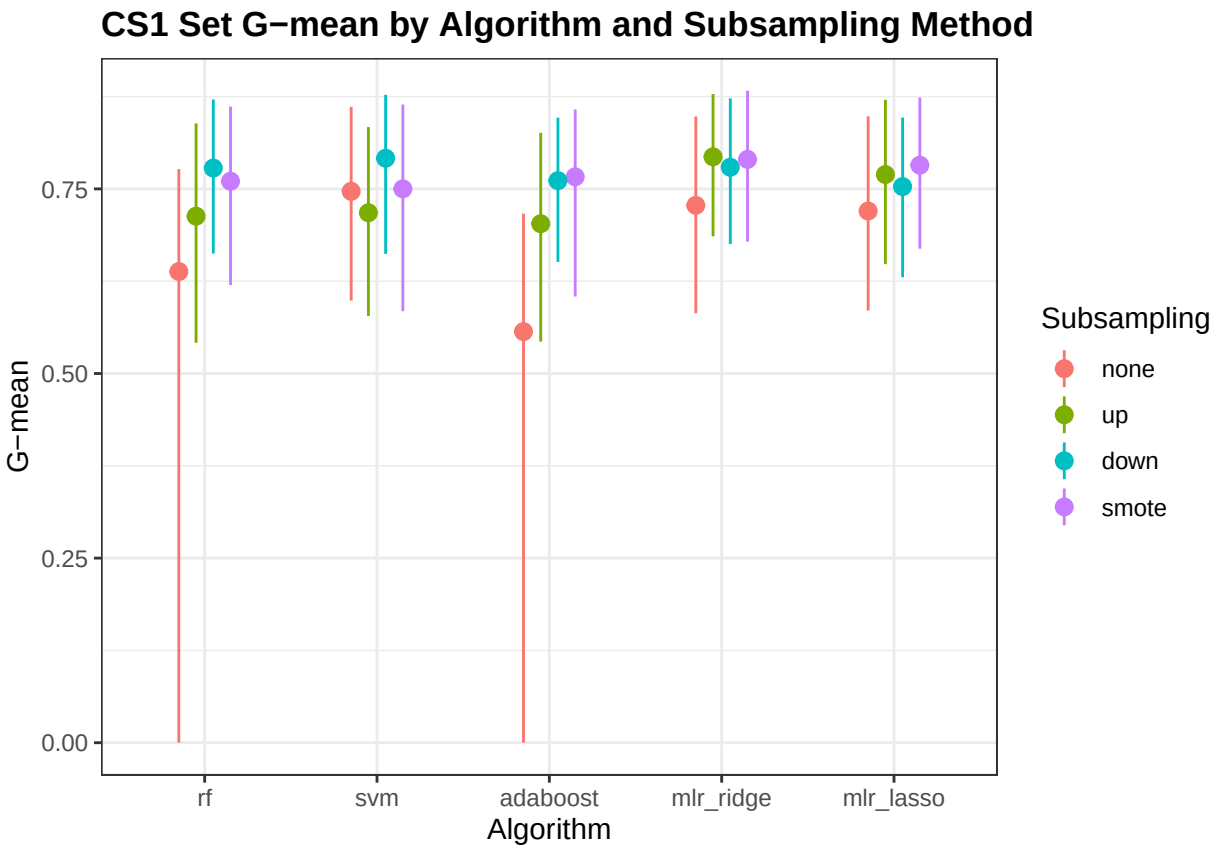


Figure 4.13: CS1 Set G-mean

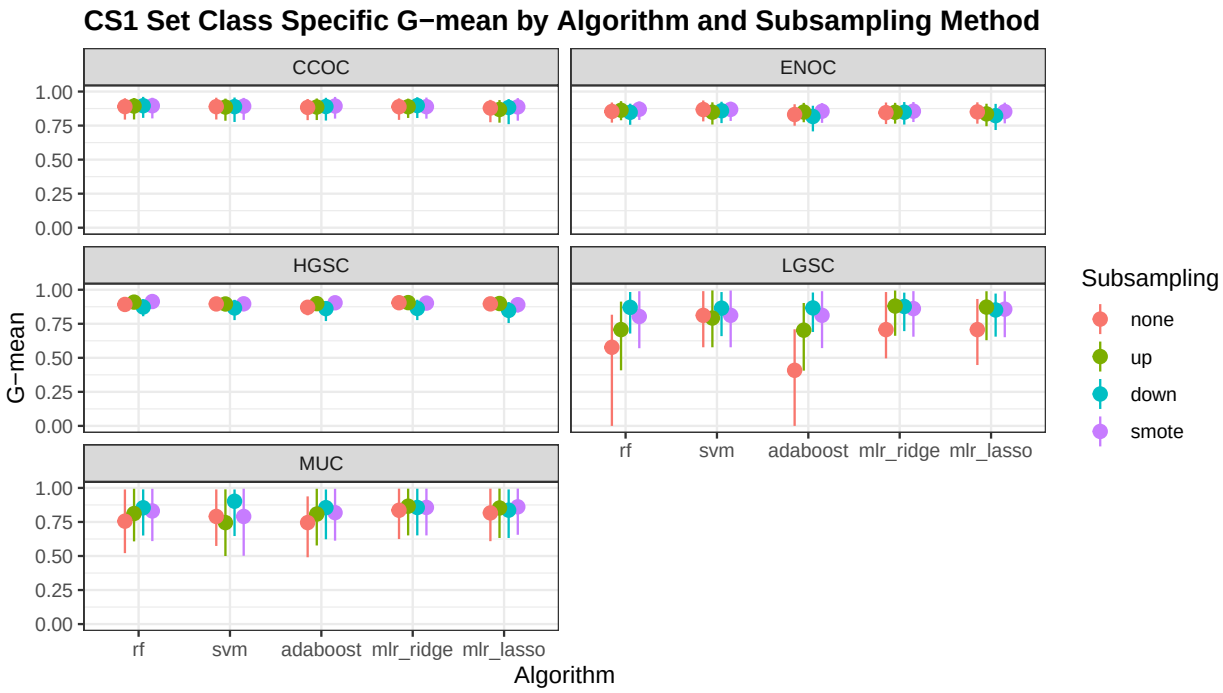


Figure 4.14: CS1 Set Class-Specific G-mean

4.3 CS2 Set

4.3.1 Accuracy

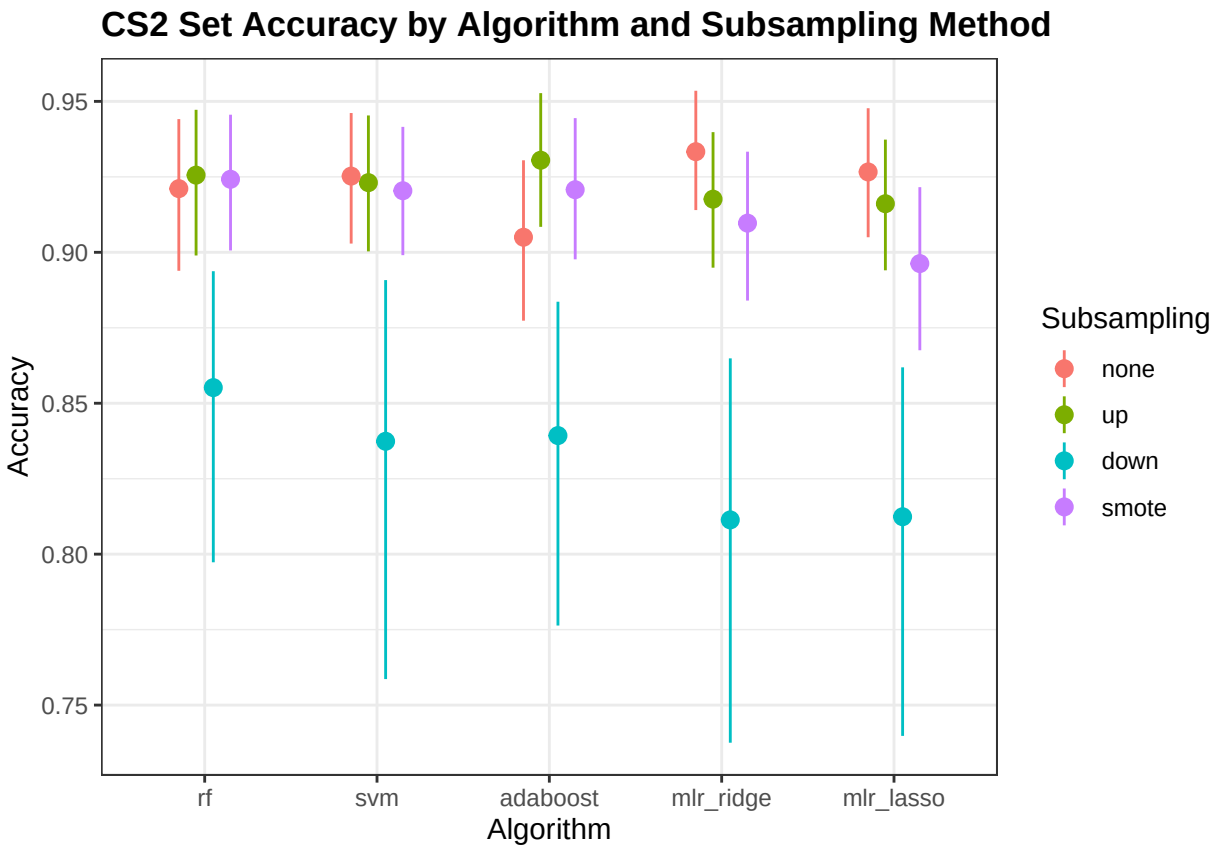


Figure 4.15: CS2 Set Accuracy

4.3.2 F1-Score

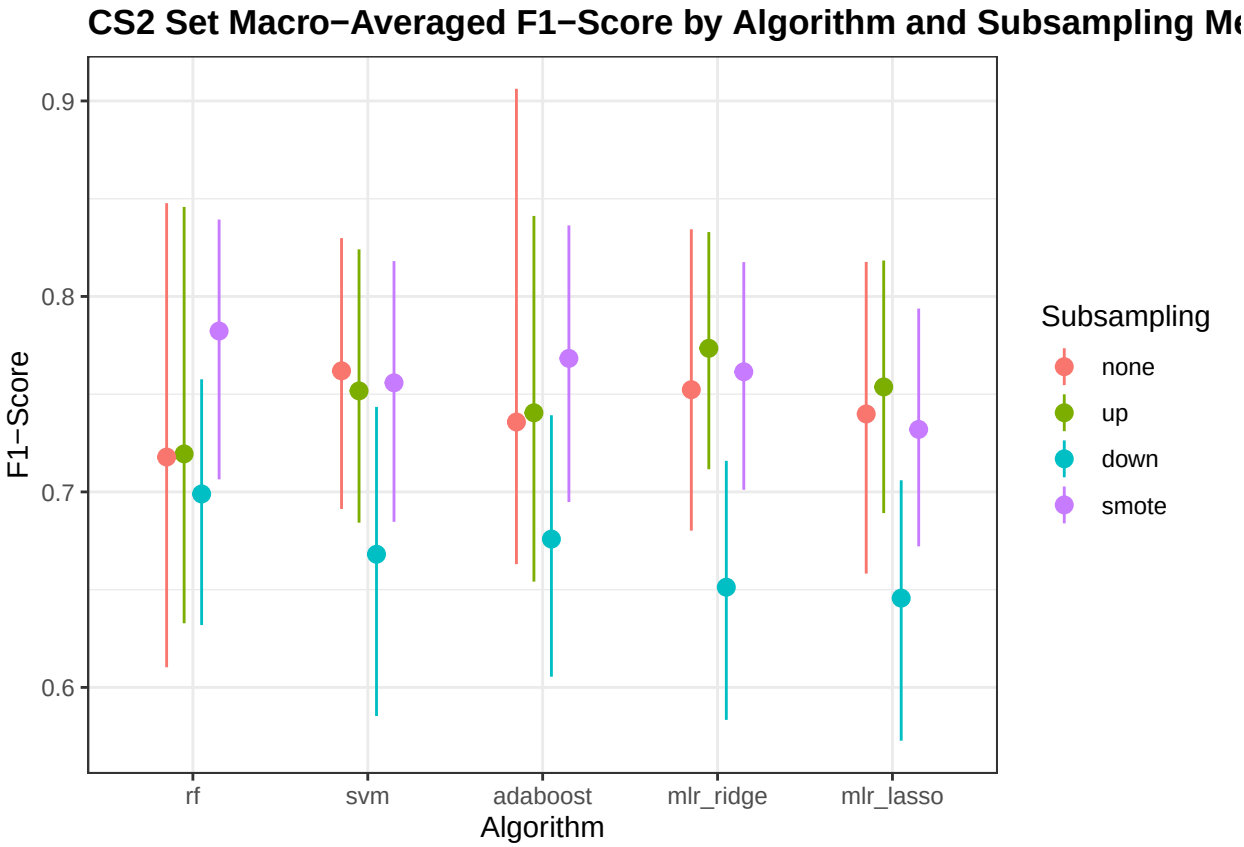


Figure 4.16: CS2 Set F1-Score

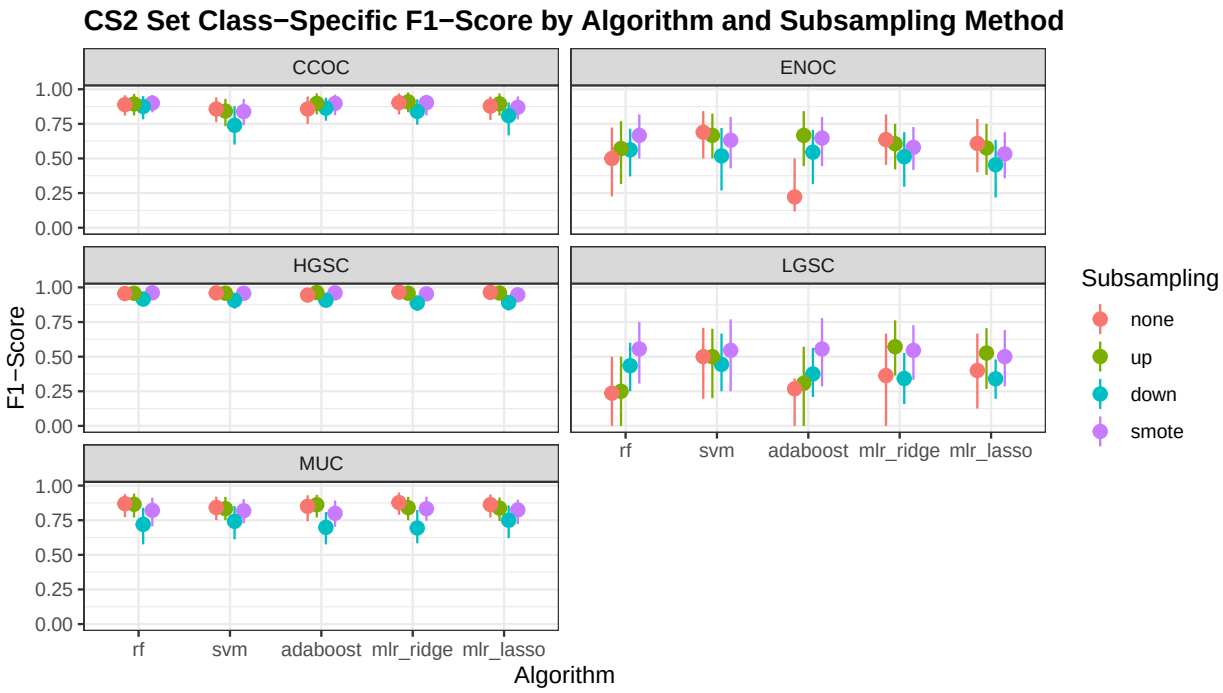


Figure 4.17: CS2 Set Class-Specific F1-Score

Table 4.3: CS2 Set Kappa by Algorithm and Subsampling Method

sampling	rf	svm	adaboost	mlr_ridge	mlr_lasso
none	0.75	0.774	0.678	**0.803**	0.788
up	0.763	0.761	0.791	0.787	0.774
down	0.669	0.629	0.641	0.599	0.594
smote	0.798	0.76	0.787	0.77	0.736

4.3.3 Kappa

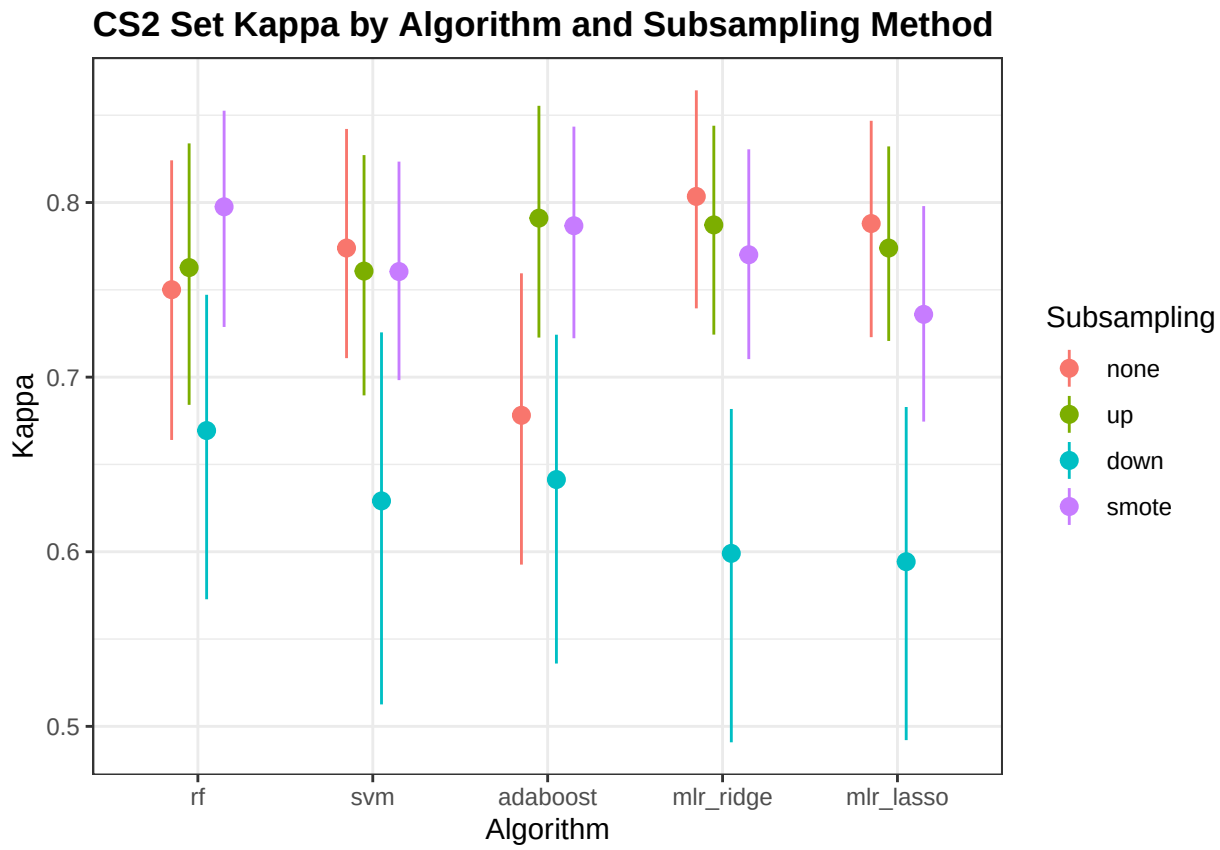


Figure 4.18: CS2 Set Kappa

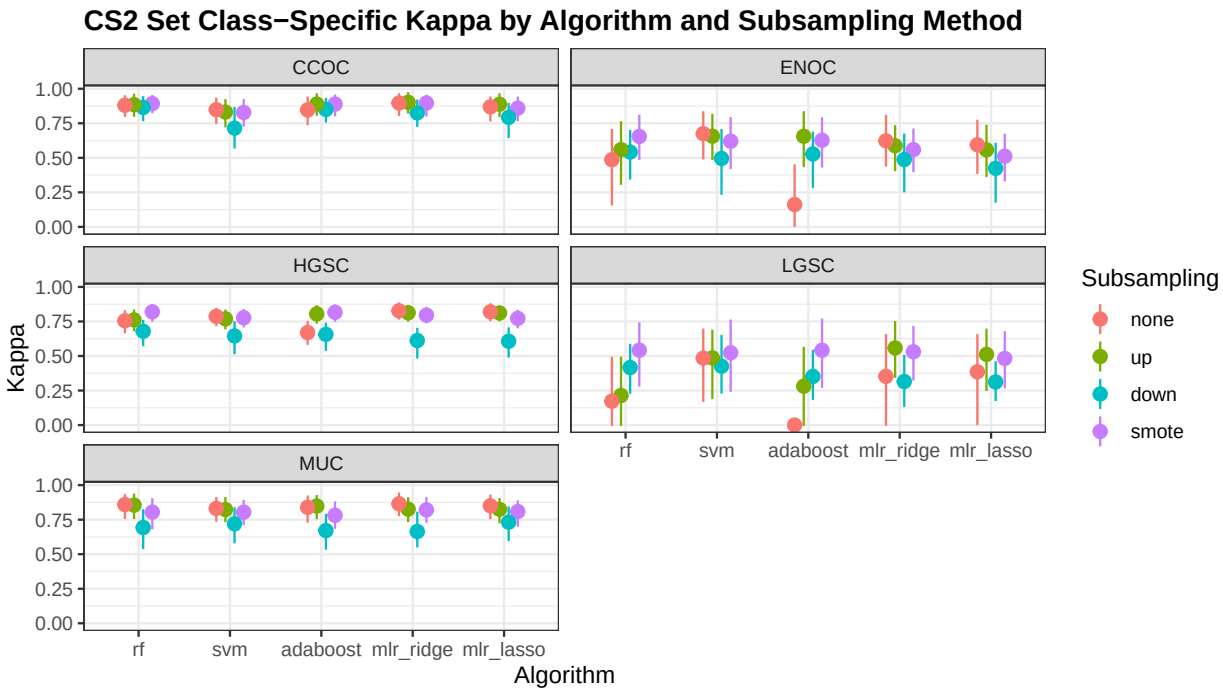


Figure 4.19: CS2 Set Class-Specific Kappa

4.3.4 G-mean

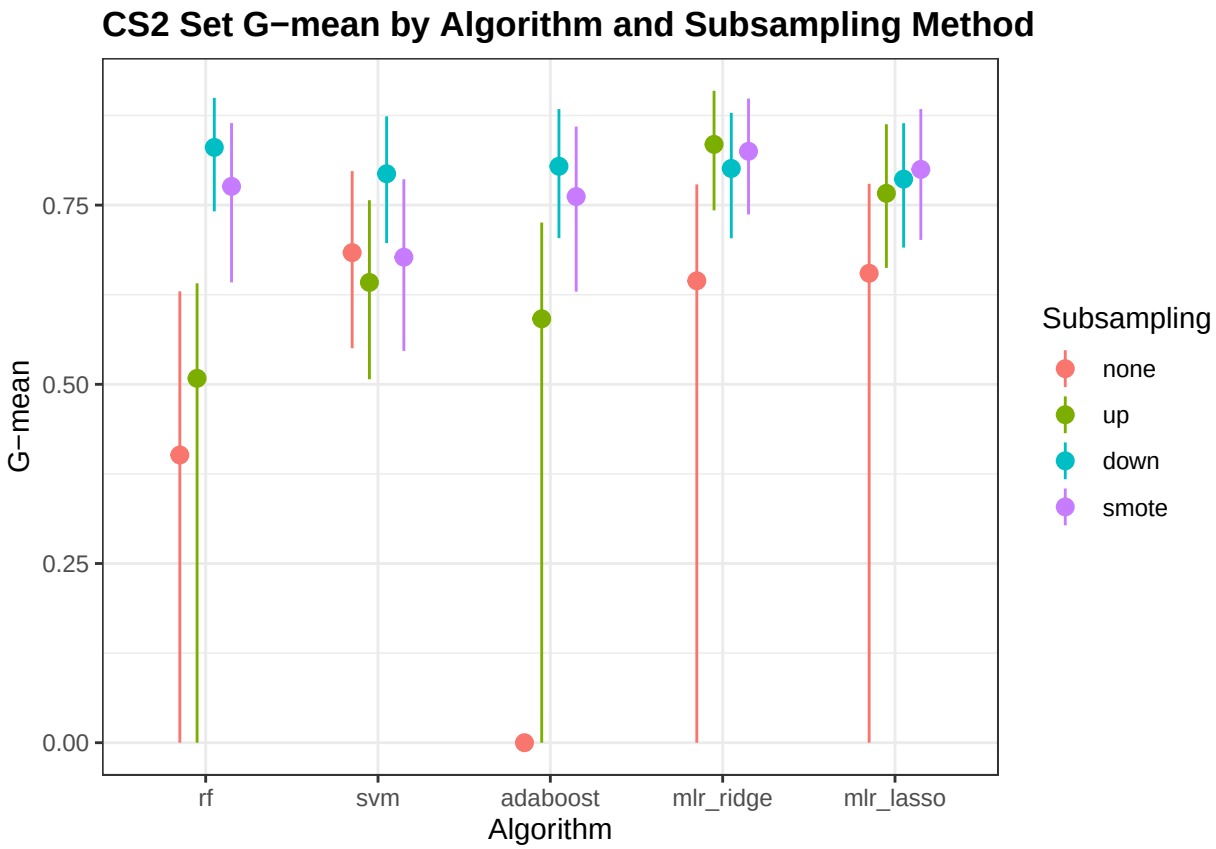


Figure 4.20: CS2 Set G-mean

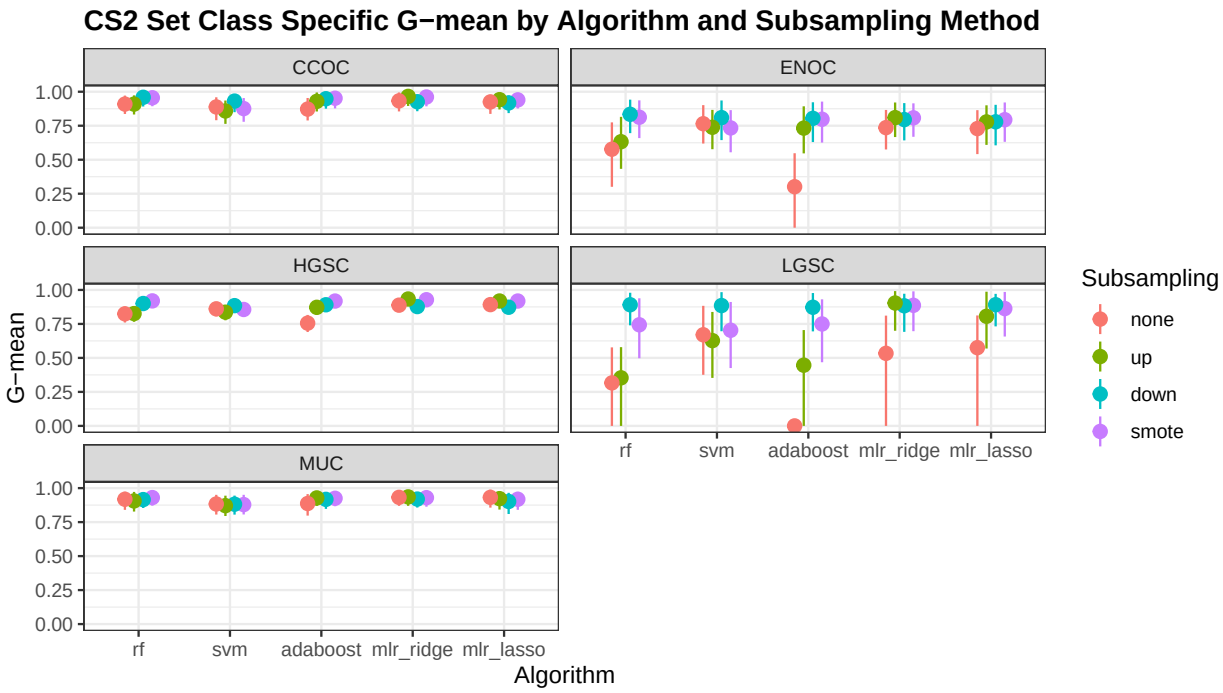


Figure 4.21: CS2 Set Class-Specific G-mean

Table 4.4: SMOTE Kappa by Algorithm and Dataset

dataset	rf	svm	adaboost	mlr_ridge	mlr_lasso
Training	**0.831**	0.811	0.81	0.766	0.758
CS1	0.763	0.759	0.751	0.755	0.736
CS2	0.798	0.76	0.787	0.77	0.736

4.4 SMOTE Kappa Summary

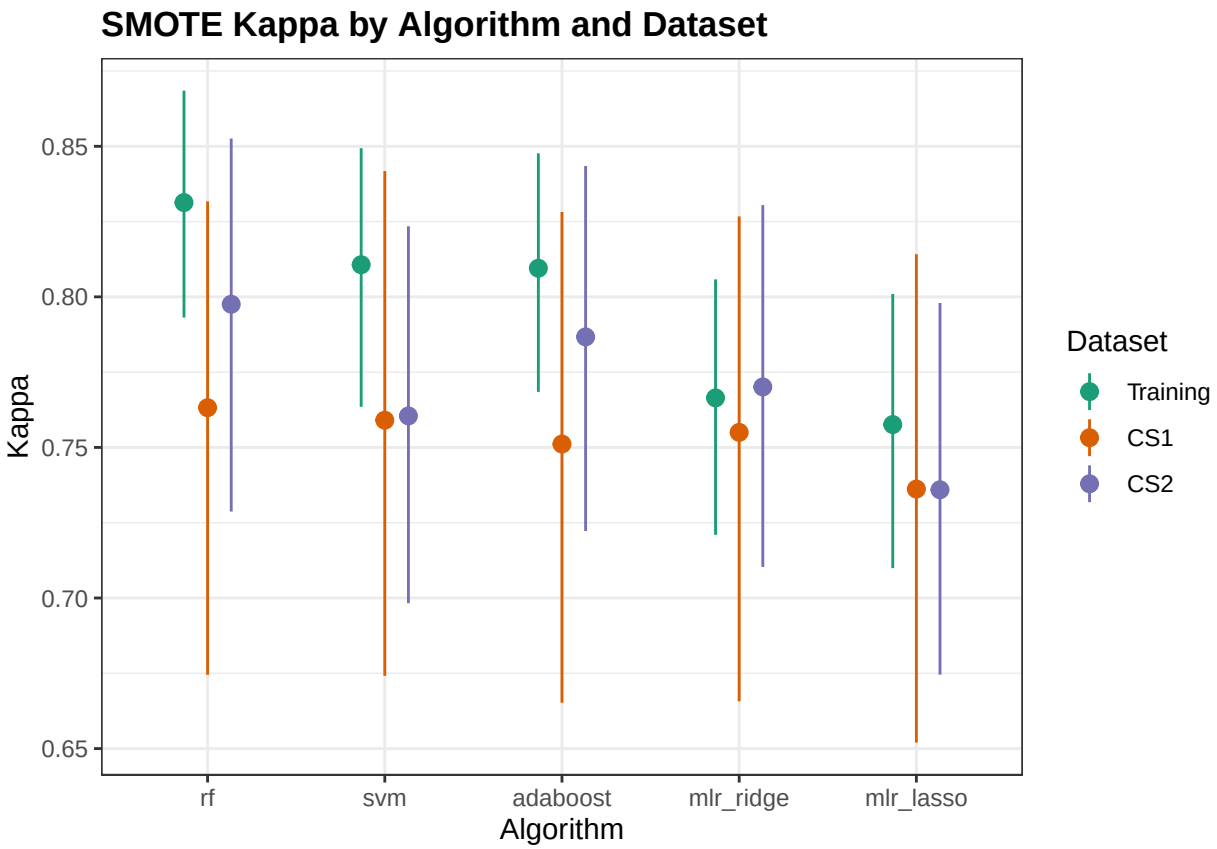


Figure 4.22: SMOTE Kappa by Algorithm and Dataset

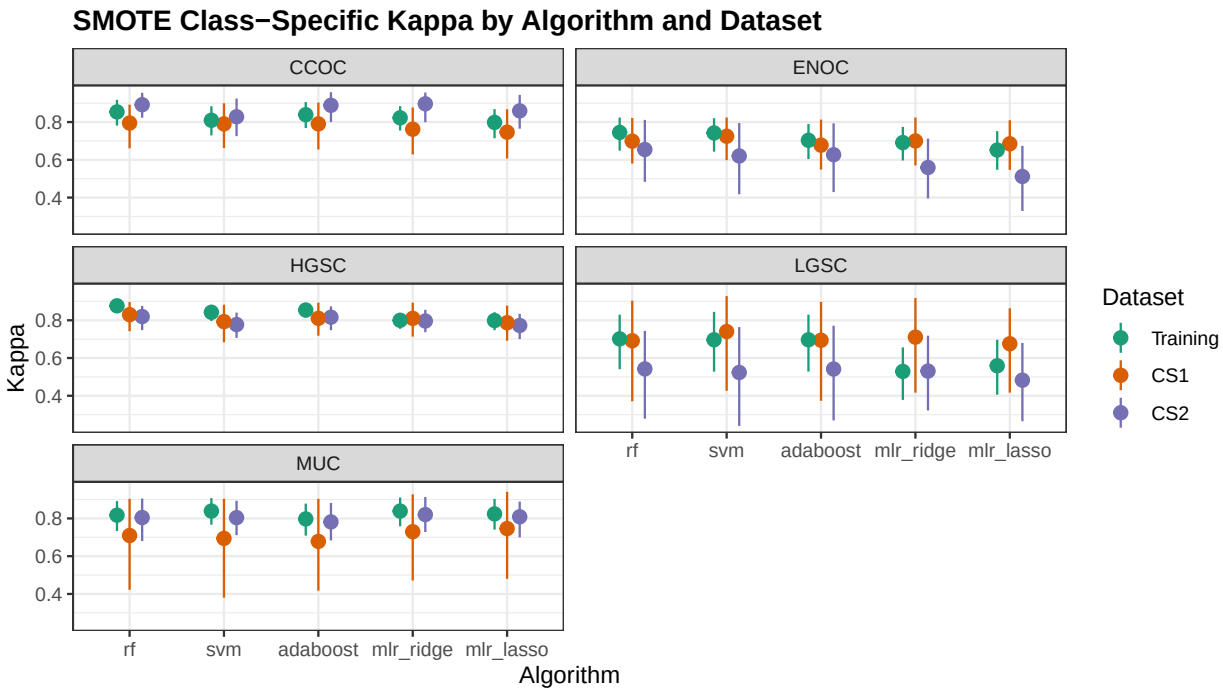


Figure 4.23: SMOTE Class-Specific Kappa by Algorithm and Dataset