Supplementary Material for Analyst

An Automated Pearson's Correlation Change Classification (APC3) approach for GC/MS metabonomic data using Total Ion Chromatograms (TIC)

Bhaskaran David Prakash[a], Kesavan Esuvaranathan[b], Paul C. Ho[a], Kishore Kumar Pasikanti[a], Eric Chun Yong, Chan[a] and Chun Wei Yap*[a]

[a] *Department of Pharmacy, National University of Singapore*
[b] *2Department of Surgery, National University Hospital, 5 Lower Kent Ridge Road, Singapore 119074.*
[*] *Department of Pharmacy, National University of Singapore, Blk S4, 18 Science Drive 4, S(117543),Tel: 065-65165971, Fax: 065-67791554, E-mail: phayapc@nus.edu.sg*

SUPPLEMENTARY MATERIAL

Table I. Mean, standard deviation (SD) and p-values in the accuracy for the various PLSR dimension reduction combinations for latent variable optimization using 10-fold cross validation(CV) versus double cross validation (CV) for a)wine set A, b)wine set B and c) wine set C. Significant higher mean and SD are highlighted in bold italics.

a)

Dimension Reduction	Average accuracy (SD) using 10-fold CV	Average accuracy (SD) using double CV	P-value
LC	***0.88(0.11)***	0.08(0.24)	1.53E-48
colAUC	***0.85(0.14)***	0.072(0.22)	3.98E-49
DDA	***0.87(0.11)***	0.076(0.23)	2.52E-50
LDA	***0.84(0.15)***	0.077(0.24)	1.60E-46
ReliefFexpRank	***0.81(0.13)***	0.078(0.24)	1.22E-47
ReliefFequalK	***0.81(0.13)***	0.078(0.24)	1.63E-47
ReliefFbestK	***0.69(0.14)***	0.063(0.19)	2.88E-47
Relief	***0.7(0.14)***	0.062(0.19)	1.36E-47
MDL	***0.84(0.12)***	0.072(0.22)	6.66E-49
Gini	***0.84(0.12)***	0.072(0.22)	6.66E-49
MyopicReliefF	***0.84(0.12)***	0.072(0.22)	6.66E-49
DKM	***0.84(0.12)***	0.072(0.22)	6.66E-49
NCA	***0.93(0.066)***	0.091(0.27)	3.91E-54
PCA1	***0.64(0.12)***	0.049(0.15)	4.05E-56
PCA2	***0.72(0.12)***	0.03(0.098)	4.45E-67
CCA1	***0.36(0)***	0.036(0.11)	2.66E-51
CCA2	***0.36(0)***	0.036(0.11)	2.66E-51
DCA	***0.88(0.066)***	0.085(0.26)	2.95E-55

b)

Dimension Reduction	Average accuracy (SD) using 10-fold CV	Average accuracy (SD) using double CV	P-value
LC	0.79(0.16)	0.79(0.15)	0.441
colAUC	0.73(0.17)	0.72(0.17)	0.326
DDA	0.78(0.17)	0.79(0.16)	0.456
LDA	0.78(0.16)	0.77(0.16)	0.429
ReliefFexpRank	0.67(0.16)	0.66(0.17)	0.565
ReliefFequalK	0.67(0.16)	0.67(0.17)	0.828
ReliefFbestK	0.59(0.14)	0.62(0.16)	0.138
Relief	0.6(0.14)	0.62(0.15)	0.156
MDL	0.73(0.18)	0.72(0.17)	0.374
Gini	0.73(0.18)	0.72(0.17)	0.374
MyopicReliefF	0.74(0.17)	0.74(0.16)	0.779
DKM	0.73(0.18)	0.72(0.17)	0.374
NCA	***0.55(0.15)***	0.048(0.15)	1.84E-42
PCA1	***0.31(0.12)***	0.06(0.18)	6.21E-21
PCA2	***0.25(0.099)***	0.022(0.075)	7.53E-32
CCA1	***0.52(0.041)***	0.045(0.14)	7.80E-58
CCA2	***0.52(0.041)***	0.045(0.14)	7.80E-58
DCA	***0.53(0.13)***	0.06(0.19)	8.12E-37

c)

Dimension Reduction	Average accuracy (SD) using 10-fold CV	Average accuracy (SD) using double CV	P-value
LC	***1(0.026)***	0.97(0.047)	3.14E-05
colAUC	***0.93(0.095)***	0.9(0.1)	0.0011
DDA	***1(0.026)***	0.97(0.047)	3.14E-05
LDA	***1(0.022)***	0.97(0.051)	3.74E-07
ReliefFexpRank	***0.95(0.1)***	0.93(0.11)	0.0436
ReliefFequalK	0.95(0.1)	0.93(0.11)	0.102
ReliefFbestK	***0.82(0.2)***	0.72(0.21)	8.14E-05
Relief	***0.79(0.21)***	0.71(0.21)	0.000207
MDL	***0.93(0.095)***	0.9(0.1)	0.0011
Gini	***0.93(0.095)***	0.9(0.1)	0.0011
MyopicReliefF	***0.93(0.095)***	0.9(0.1)	0.0011
DKM	***0.93(0.095)***	0.9(0.1)	0.0011
NCA	***0.94(0.073)***	0.094(0.28)	1.29E-46
PCA1	***0.4(0.14)***	0.031(0.096)	3.25E-39
PCA2	***0.68(0.15)***	0.058(0.18)	1.83E-50
CCA1	***0.3(0)***	0.03(0.09)	2.66E-51
CCA2	***0.3(0)***	0.03(0.09)	2.66E-51
DCA	***0.95(0.059)***	0.094(0.28)	1.53E-50

Table II. Mean, standard deviation (SD) and p-values in the AUC for the various PLSR dimension reduction combinations for latent variable optimization using 10-fold cross validation(CV) versus double cross validation (CV) for a)wine set A, b)wine set B and c) wine set C. Significant higher mean and SD are highlighted in bold italics.

a)

Dimension Reduction	Average AUC (SD) using 10-fold CV	Average AUC (SD) using double CV	P-value
LC	***0.86(0.11)***	0.081(0.25)	1.18E-47
colAUC	***0.84(0.14)***	0.071(0.22)	6.66E-49
DDA	***0.85(0.11)***	0.078(0.24)	4.18E-49
LDA	***0.83(0.15)***	0.077(0.24)	3.81E-46
ReliefFexpRank	***0.81(0.13)***	0.079(0.24)	1.73E-47
ReliefFequalK	***0.8(0.13)***	0.079(0.24)	2.16E-47
ReliefFbestK	***0.68(0.14)***	0.063(0.2)	4.40E-46
Relief	***0.69(0.14)***	0.063(0.19)	8.80E-47
MDL	***0.82(0.13)***	0.071(0.22)	1.48E-47
Gini	***0.82(0.13)***	0.071(0.22)	1.48E-47
MyopicReliefF	***0.82(0.13)***	0.071(0.22)	1.48E-47
DKM	***0.82(0.13)***	0.071(0.22)	1.48E-47
NCA	***0.91(0.083)***	0.089(0.27)	1.83E-54
PCA1	***0.62(0.13)***	0.039(0.12)	8.82E-60
PCA2	***0.71(0.12)***	0.033(0.11)	8.19E-67
CCA1	***0.5(0)***	0.05(0.15)	2.66E-51
CCA2	***0.5(0)***	0.05(0.15)	2.66E-51
DCA	***0.85(0.082)***	0.084(0.25)	8.86E-54

b)

Dimension Reduction	Average AUC (SD) using 10-fold CV	Average AUC (SD) using double CV	P-value
LC	0.78(0.15)	0.79(0.14)	0.344
colAUC	0.72(0.16)	0.71(0.16)	0.255
DDA	0.78(0.16)	0.79(0.15)	0.457
LDA	0.77(0.15)	0.77(0.15)	0.434
ReliefFexpRank	0.68(0.16)	0.66(0.17)	0.373
ReliefFequalK	0.67(0.16)	0.66(0.18)	0.611
ReliefFbestK	0.6(0.14)	0.62(0.16)	0.197
Relief	0.6(0.14)	0.62(0.15)	0.223
MDL	0.72(0.17)	0.71(0.17)	0.348
Gini	0.72(0.17)	0.71(0.17)	0.348
MyopicReliefF	0.74(0.17)	0.73(0.16)	0.677
DKM	0.72(0.17)	0.71(0.17)	0.348
NCA	***0.55(0.15)***	0.047(0.15)	3.57E-44
PCA1	***0.3(0.12)***	0.057(0.17)	8.58E-22
PCA2	***0.25(0.098)***	0.021(0.071)	7.58E-32
CCA1	***0.5(0)***	0.05(0.15)	2.66E-51
CCA2	***0.5(0)***	0.05(0.15)	2.66E-51
DCA	***0.54(0.13)***	0.06(0.18)	1.46E-37

c)

Dimension Reduction	Average AUC (SD) using 10-fold CV	Average AUC (SD) using double CV	P-value
LC	***1(0.019)***	0.96(0.077)	1.30E-06
colAUC	***0.92(0.11)***	0.89(0.13)	0.000188
DDA	***1(0.019)***	0.96(0.077)	1.30E-06
LDA	***1(0.016)***	0.95(0.08)	9.45E-08
ReliefFexpRank	***0.94(0.12)***	0.91(0.14)	0.00825
ReliefFequalK	***0.94(0.12)***	0.91(0.14)	0.0213
ReliefFbestK	***0.8(0.22)***	0.68(0.21)	1.03E-06
Relief	***0.78(0.21)***	0.68(0.21)	1.11E-06
MDL	***0.92(0.11)***	0.89(0.13)	0.000188
Gini	***0.92(0.11)***	0.89(0.13)	0.000188
MyopicReliefF	***0.92(0.11)***	0.89(0.13)	0.000188
DKM	***0.92(0.11)***	0.89(0.13)	0.000188
NCA	***0.9(0.11)***	0.09(0.27)	4.31E-44
PCA1	***0.41(0.15)***	0.024(0.076)	3.56E-39
PCA2	***0.71(0.14)***	0.062(0.19)	9.89E-49
CCA1	***0.5(0)***	0.05(0.15)	2.66E-51
CCA2	***0.5(0)***	0.05(0.15)	2.66E-51
DCA	***0.93(0.077)***	0.09(0.27)	4.51E-50

Table III. Overall average accuracy and average AUC for each transformation/variable and classification combination using the wine testing sets. Values are expressed in (average accuracy ± standard deviation, average AUC ± standard deviation). Top 2 approaches values for average accuracy and average AUC highlighted in bold italics.

		Transformation					
		NCA	PCA_1	PCA_2	CCA_1	CCA_2	DCA
Classification	PLSR	0.81±0.22 0.79±0.10	0.45±0.17 0.44±0.16	0.55±0.26 0.56±0.27	0.39±0.11 0.50±0.00	0.39±0.11 0.50±0.00	0.78±0.23 0.77±0.20
	NB	0.69±0.08 0.71±0.06	0.62±0.16 0.60±0.14	0.59±0.09 0.59±0.13	0.68±0.14 0.67±0.13	0.68±0.14 0.67±0.13	0.78±0.16 0.78±0.15
	locLDA	0.57±0.10 0.57±0.10	0.47±0.07 0.47±0.05	0.51±0.08 0.51±0.08	0.47±0.05 0.51±0.01	0.46±0.05 0.49±0.01	0.77±0.20 0.77±0.19

		Variable Selection					
		LC	colAUC	DDA	LDA	ReliefFexpRank	ReliefFequalK
Classification	PLSR	0.89±0.10 0.88±0.11	0.84±0.10 0.83±0.10	0.88±0.11 0.88±0.11	0.87±0.11 0.87±0.12	0.81±0.14 0.81±0.13	0.81±0.14 0.80±0.13
	NB	***0.89±0.10** ***0.89±0.10***	0.84±0.09 0.83±0.09	***0.90±0.09*** ***0.89±0.10***	0.88±0.10 0.88±0.10	0.81±0.12 0.80±0.11	0.81±0.12 0.80±0.11
	locLDA	0.88±0.10 0.88±0.10	0.84±0.10 0.83±0.09	0.88±0.10 0.89±0.10	0.87±0.11 0.88±0.10	0.81±0.14 0.80±0.11	0.81±0.14 0.80±0.11

		Variable Selection					
		ReliefFbestK	Relief	MDL	Gini	MyopicReliefF	DKM
Classification	PLSR	0.70±0.11 0.69±0.10	0.70±0.10 0.69±0.09	0.83±0.10 0.82±0.10	0.83±0.10 0.82±0.10	0.84±0.09 0.83±0.09	0.83±0.10 0.82±0.10
	NB	0.76±0.07 0.75±0.06	0.75±0.06 0.75±0.05	0.85±0.09 0.84±0.09	0.85±0.09 0.85±0.09	0.85±0.08 0.85±0.09	0.85±0.09 0.85±0.09
	locLDA	0.70±0.10 0.75±0.06	0.69±0.09 0.75±0.05	0.83±0.11 0.85±0.09	0.83±0.11 0.85±0.09	0.84±0.09 0.85±0.09	0.83±0.11 0.85±0.09

*LC-NB had an average accuracy of 0.892 which was higher than that of LC-PLSR's 0.886

Table IV. Variable length after transformation

		Wine Set A	Wine Set B	Wine Set C
Transformation	NCA	2700		
	*PCA_1	26	22	25
	*PCA_2			
	*CCA_1			
	*CCA_2			
	**DCA	4		

* (Number of samples) – 1

** Only finds 4 axes according to the original implementation in M. O. Hill and H. G. Gauch, Vegetatio, 1980, 42, 47-58.

Table V. Mean number of latent variable (SD) after transformation and PLSR latent variable optimization

		Wine Set A	Wine Set B	Wine Set C
Transformation	NCA	2.75(0.60)	3.01(0.89)	2.63(0.56)
	PCA_1	1(0)	1(0)	1(0)
	PCA_2	1(0)	1(0)	1(0)
	CCA_1	1(0)	1(0)	1(0)
	CCA_2	1(0)	1(0)	1(0)
	DCA	1.56(0.77)	1.76(0.77)	1.14(0.495)

Table VI. Mean number of latent variable (SD) for training model using PLSR after variable selection

		Wine Set A	Wine Set B	Wine Set C
Variable Selection	LC	4.63(2.63)	2.60(1.38)	2(0)
	colAUC	4.05(2.18)	3.11(1.63)	2.77(0.77)
	DDA	4.73(2.61)	2.61(1.46)	2(0)
	LDA	4.05(1.89)	2.74(1.54)	2.01(0.10)
	ReliefFexpRank	6.12(2.88)	5.38(2.13)	2.57(0.96)
	ReliefFequalK	6.16(2.91)	5.41(2.11)	2.56(0.96)
	ReliefFbestK	7.00(2.70)	6.07(2.43)	5.71(2.48)
	Relief	7.10(2.76)	6.04(2.39)	5.75(2.55)
	MDL	5.46(2.60)	3.57(2.34)	2.77(0.77)
	Gini	5.46(2.60)	3.56(2.31)	2.77(0.77)
	MyopicReliefF	5.46(2.60)	3.21(1.73)	2.77(0.77)
	DKM	5.46(2.60)	3.56(2.31)	2.77(0.77)

Table VII. Mean variable length (SD) for training model using NB after variable selection

		Wine Set A	Wine Set B	Wine Set C
Variable Selection	LC	10.60(9.75)	3.13(3.50)	2(0)
	colAUC	6.97(7.22)	5.05(8.16)	2.84(1.08)
	DDA	11.70(10.60)	3.15(3.44)	2(0)
	LDA	6.56(5.32)	4.42(5.58)	2.01(0.10)
	ReliefFexpRank	9.67(9.34)	6.67(7.51)	2.98(1.90)
	ReliefFequalK	9.83(10.20)	6.58(7.34)	2.92(1.86)
	ReliefFbestK	15.90(15.10)	8.58(11.20)	11.70(10.90)
	Relief	16.30(15.60)	8.16(10.80)	11.60(11.30)
	MDL	14.10(11.30)	5.48(8.93)	2.84(1.08)
	Gini	14.10(11.30)	5.48(8.93)	2.84(1.08)
	MyopicReliefF	13.80(11.00)	4.29(7.47)	2.84(1.08)
	DKM	14.10(11.30)	5.48(8.93)	2.84(1.08)

Table VIII. Mean variable length (SD) for training model using locLDA after variable selection

		Wine Set A	Wine Set B	Wine Set C
Variable Selection	LC	3.66(1.69)	2.46(1.18)	2(0)
	colAUC	3.87(2.06)	2.80(1.29)	2.92(0.99)
	DDA	3.66(1.69)	2.49(1.29)	2(0)
	LDA	3.49(1.40)	2.59(1.40)	2.01(0.10)
	ReliefFexpRank	5.06(2.42)	4.48(1.77)	2.27(0.53)
	ReliefFequalK	5.05(2.49)	4.45(1.79)	2.25(0.52)
	ReliefFbestK	6.08(2.51)	5.16(2.25)	5.18(2.34)
	Relief	6.16(2.57)	5.16(2.20)	5.36(2.44)
	MDL	4.75(2.10)	3.13(1.74)	2.92(0.99)
	Gini	4.75(2.10)	3.13(1.74)	2.92(0.99)
	MyopicReliefF	4.75(2.10)	2.83(1.32)	2.92(0.99)
	DKM	4.75(2.10)	3.13(1.74)	2.92(0.99)

Table IX. Mean variable length (SD) for training model using urine sample splits after variable selection for both DDA-NB and LC-NB

	Urine Sample Split					
	A	B	C	D	E	F
DDA-NB	19.4(12.8)	12.5(11.9)	7.42(6.49)	5.49(4.17)	5.75(6.47)	5.01(3.23)
LC-NB	18.5(12.6)	11.2(11.8)	6.39(7.54)	3.83(4.17)	3.83(6.35)	2.81(3.17)

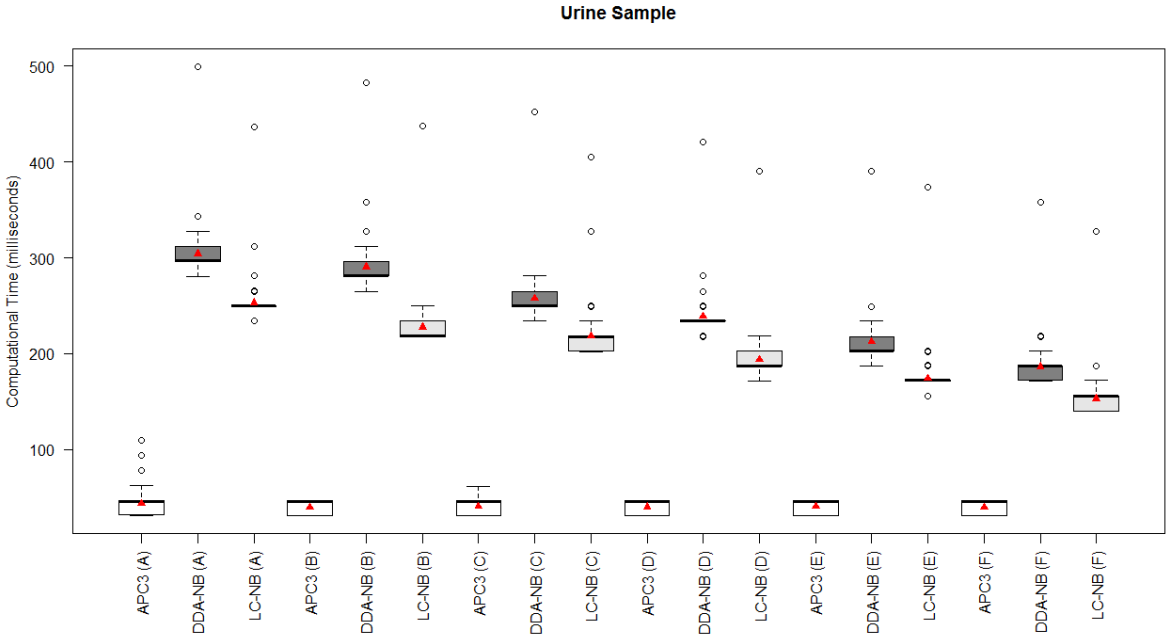

Fig Ia. Boxplot of computational time (both training and testing phases) across the 100 sampling for urine sample splits using APC3, DDA-NB and LC-NB. The red triangle represents the mean value of each boxplot.

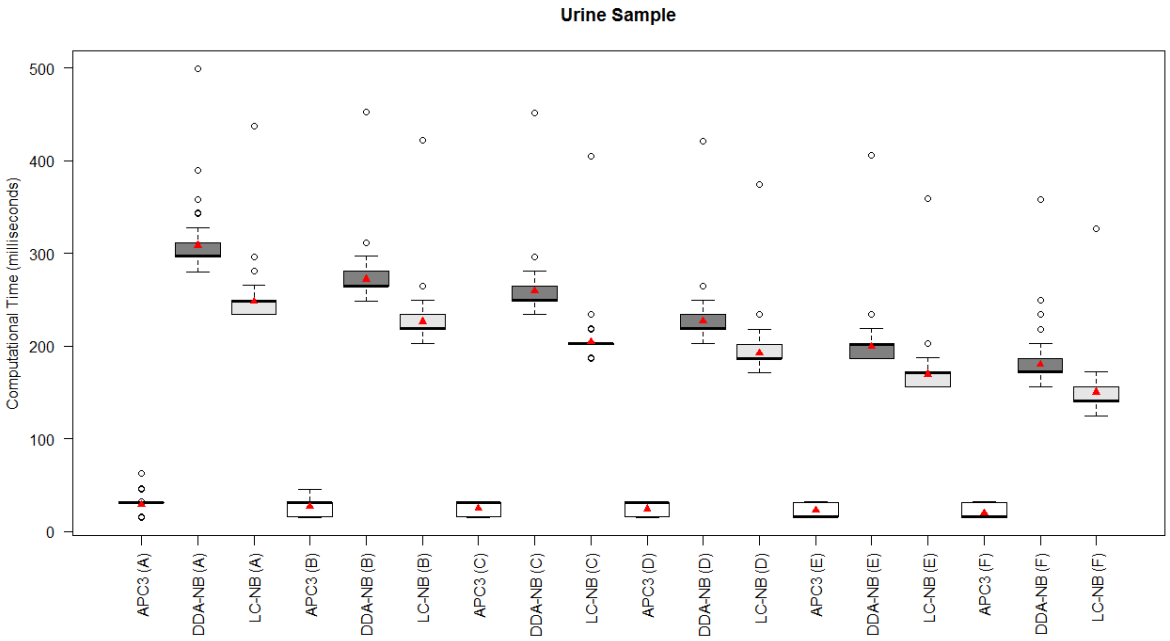

Fig Ib. Boxplot of computational time (only training phase) across the 100 sampling for urine sample splits using APC3, DDA-NB and LC-NB. The red triangle represents the mean value of each boxplot.

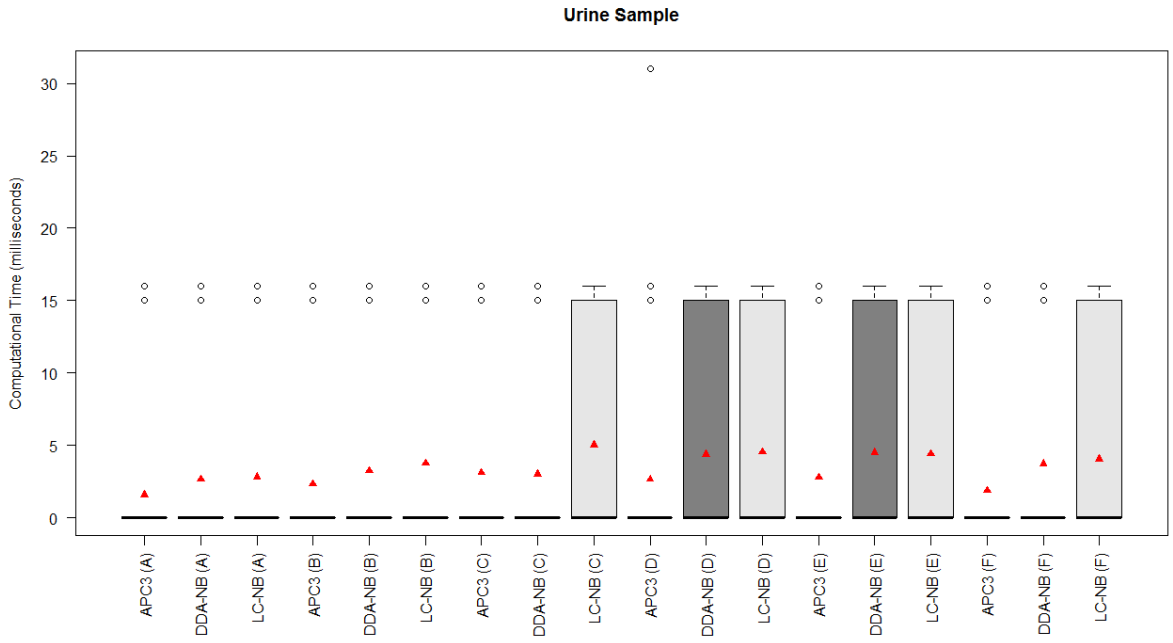

Fig Ic. Boxplot of computational time (only testing phase) across the 100 sampling for urine sample splits using APC3, DDA-NB and LC-NB. The red triangle represents the mean value of each boxplot.