Figures and tables for the paper

Tusharkanti Ghosh, Daisy Philtron, Weiming Zhang, Katerina Kechris and Debashis Ghosh

September 24, 2020

1 Supplementary results

We applied the MaRR procedure to each of the 3 MS-Metabolomics data sets. The reproducible signals were identified for all 3 layers (batch, spike-in and replicates) in Tech data set; 2 layers (biological subjects and replicates) in BioTech data set and a single layer (biological replicates) in Bio data set.

We pre-processed the data using the MSPrep software. The main two steps: i) Filtering missing metabolites (metabolites were removed if they were missing more than 80% of the samples); ii) Imputing missing values- Bayesian Principal Component Analysis (BPCA) and kNN; iii) Normalization- median, quantile and RUV.

The main purpose of applying the MaRR procedure to the 3 MS-Metabolomics processed (6 combinations of pre-processing steps: 3 normalization methods and 2 missing value imputations) data sets was to identify the pre-processing step combination that produce the maximum reproducible metabolites and sample pairs. The MS-Metabolomics data set (pre-processed using a unique method combination of missing value imputation and normalization, i.e., BPCA and RUV) generated the maximum reproducible signals using the MaRR procedure. Summaries of the best results (Missing value of imputation: BPCA and Normalization: RUV) in terms of maximum reproducibility are presented in Table 7.2.

2 Imputation- kNN and Normalization- median

2.1 Tech data

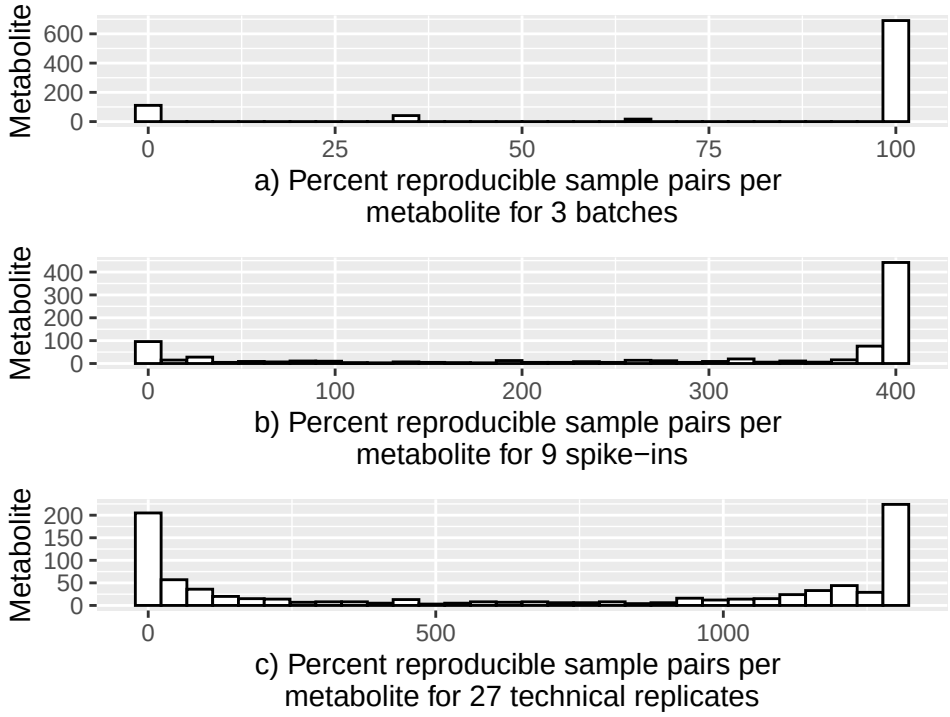

Figure 1: Reproducibility of sample pairs per metabolite.

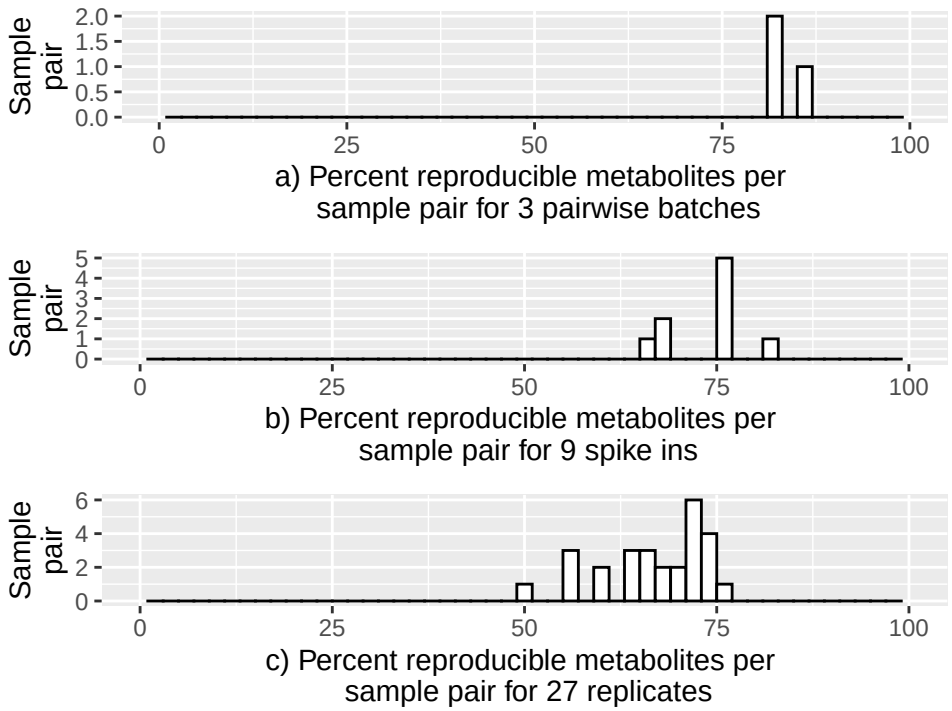

Figure 3: Reproducibility of metabolites per sample pair.

2.2 BioTech and Bio data

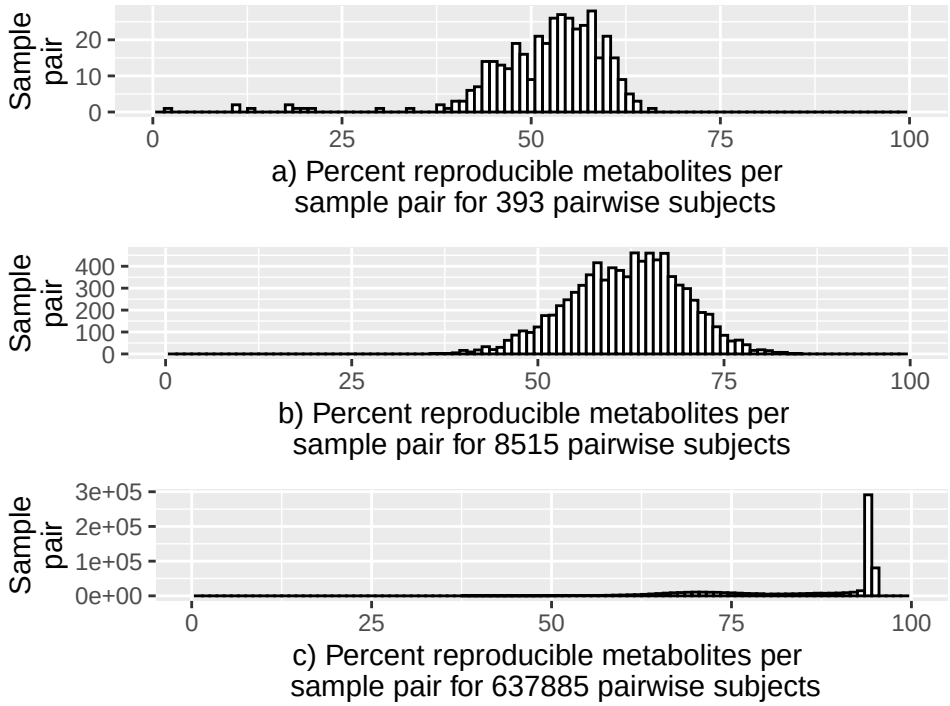

Figure 2: Reproducibility of metabolites per sample pair.

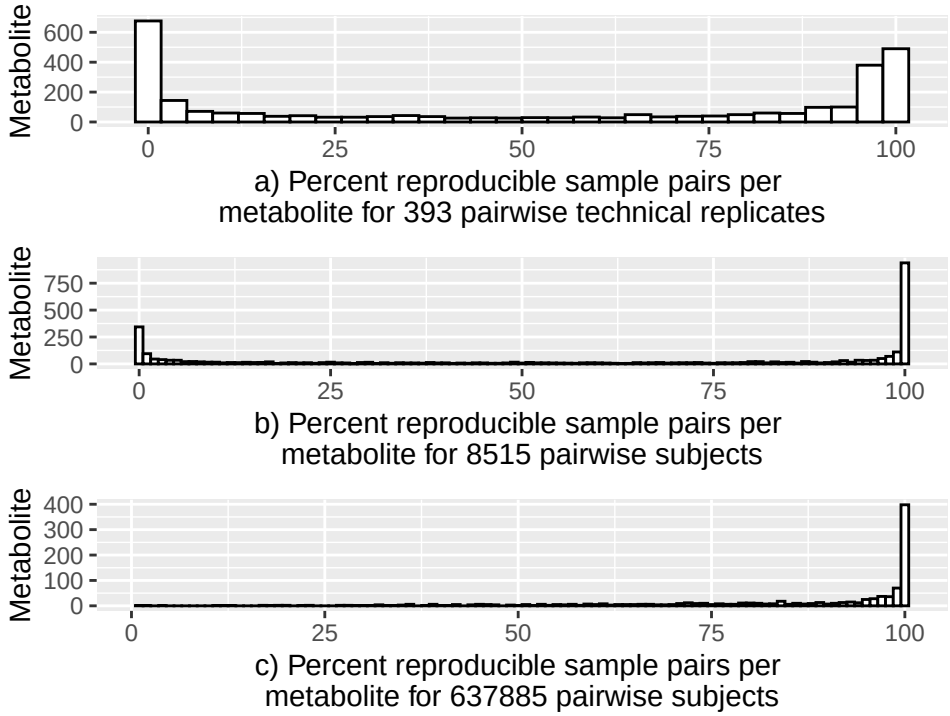

Figure 4: Reproducibility of sample pairs per metabolite.

Data set		Percentage of reproducible metabolites per sample pair			Percentage of reproducible sample pairs per metabolite		
		$> 70\%$	$> 80\%$	$> 90\%$	$> 70\%$	$> 80\%$	$> 90\%$
Tech	Operator layer	100.00	100.00	0.00	80.35	80.35	80.35
	Spike-in layer	66.67	11.11	0.00	81.40	81.16	80.12
	Replicate layer	40.74	0.00	0.00	69.53	68.37	67.21
BioTech	biological layer	14.14	0.47	0.00	56.01	52.41	47.06
	technical layer	0.00	0.00	0.00	46.05	41.89	35.45
Bio	biological layer	87.62	74.39	64.33	83.22	75.08	65.23

Table 1: Summary of reproducible metabolites per sample pair and reproducible sample pairs per metabolite for Tech, BioTech and Bio data sets. The columns of proportion of reproducible metabolites per sample pair $> 70/80/90\%$ indicate that proportion of sample pairs of replicate experiments with reproducible metabolites per sample pair greater than $> 70/80/90\%$. The columns of proportion of reproducible sample pairs per metabolite $> 70/80/90\%$ indicate that proportion of metabolites with reproducible sample pairs per metabolite greater than $> 70/80/90\%$.

3 Imputation- kNN and Normalization- quantile

3.1 Tech data

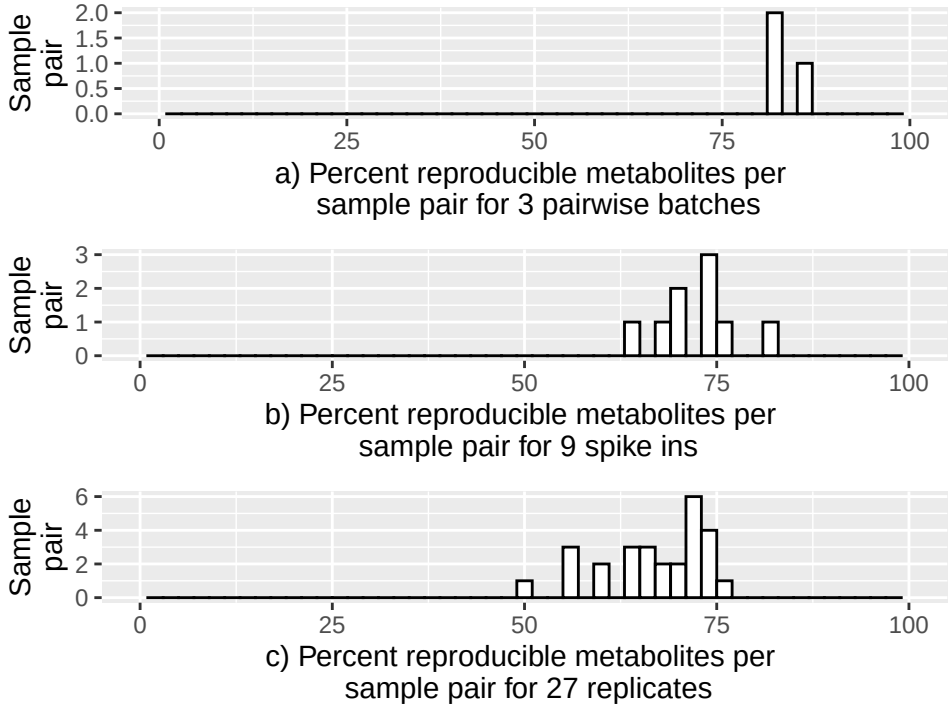

Figure 5: Reproducibility of metabolites per sample pair.

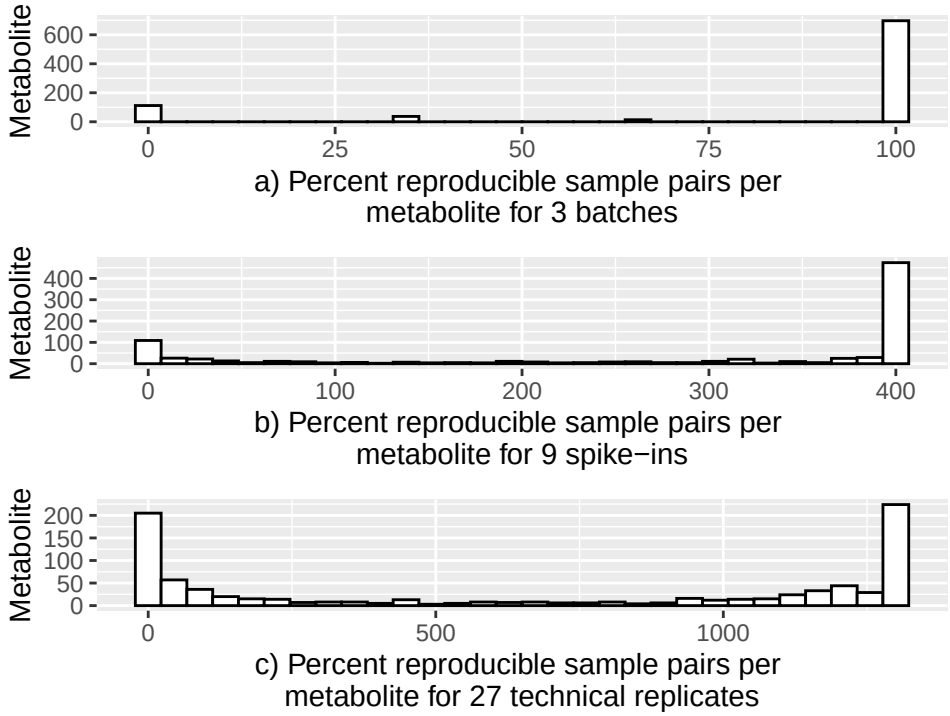

Figure 6: Reproducibility of sample pairs per metabolite.

3.2 BioTech and Bio data

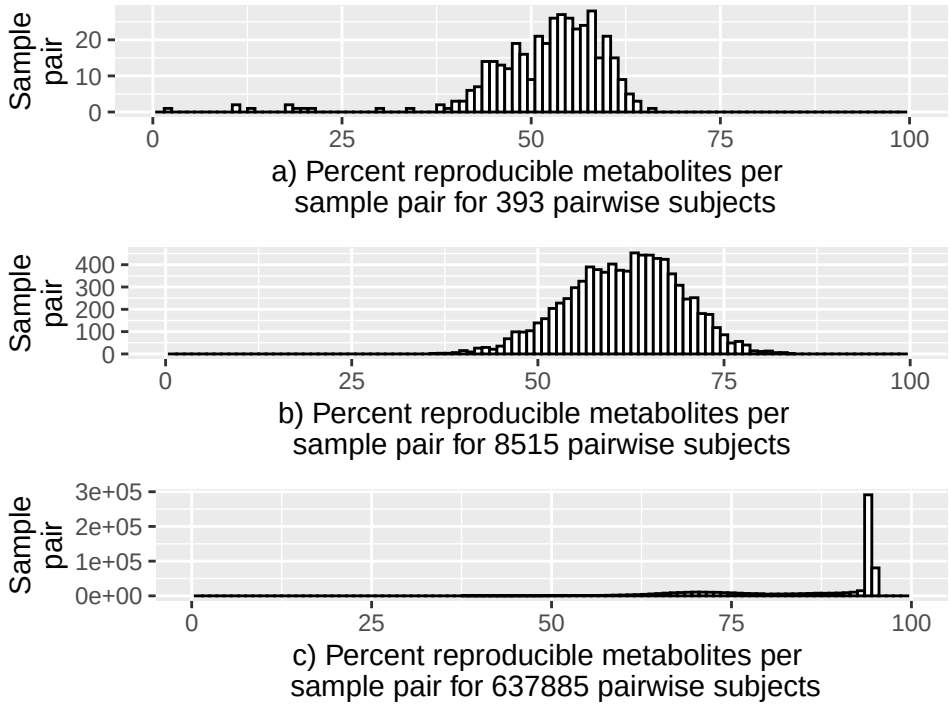

Figure 7: Reproducibility of metabolites per sample pair.

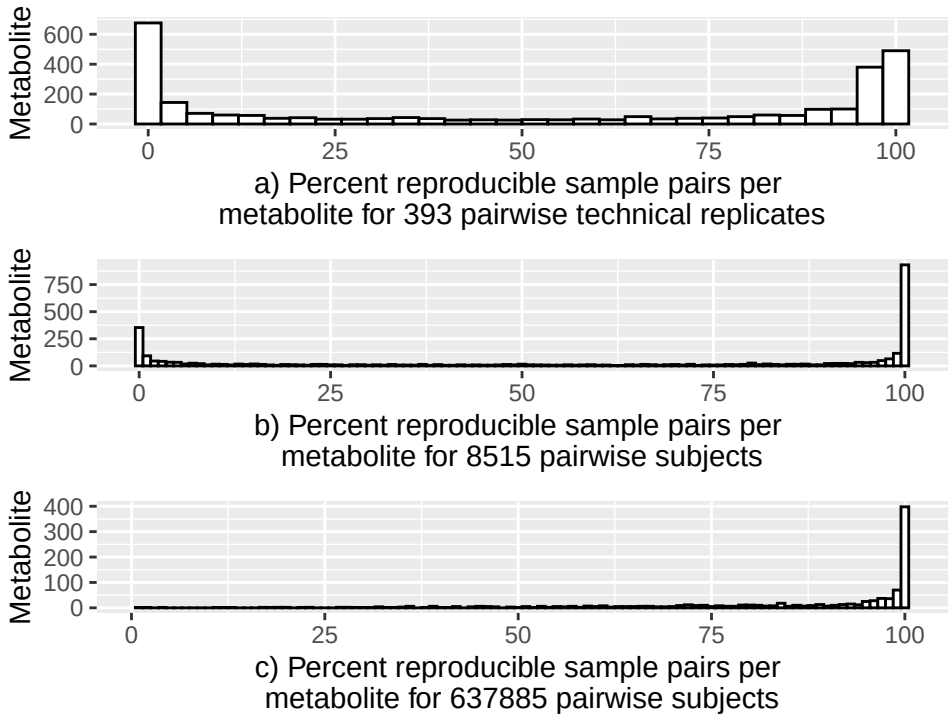

Figure 8: Reproducibility of sample pairs per metabolite.

Data set		Percentage of reproducible metabolites per sample pair			Percentage of reproducible sample pairs per metabolite		
		$> 70\%$	$> 80\%$	$> 90\%$	$> 70\%$	$> 80\%$	$> 90\%$
Tech	Operator layer	100.00	100.00	0.00	81.05	81.05	81.05
	Spike-in layer	55.56	11.11	0.00	78.37	77.79	77.33
	Replicate layer	40.74	0.00	0.00	69.53	68.37	67.21
BioTech	biological layer	13.39	0.36	0.00	55.66	51.75	46.89
	technical layer	0.00	0.00	0.00	46.05	41.89	35.45
Bio	biological layer	87.62	74.39	64.33	83.22	75.08	65.23

Table 2: Summary of reproducible metabolites per sample pair and reproducible sample pairs per metabolite for Tech, BioTech and Bio data sets. The columns of proportion of reproducible metabolites per sample pair $> 70/80/90\%$ indicate that proportion of sample pairs of replicate experiments with reproducible metabolites per sample pair greater than $> 70/80/90\%$. The columns of proportion of reproducible sample pairs per metabolite $> 70/80/90\%$ indicate that proportion of metabolites with reproducible sample pairs per metabolite greater than $> 70/80/90\%$.

4 Imputation- kNN and Normalization- RUV

4.1 Tech data

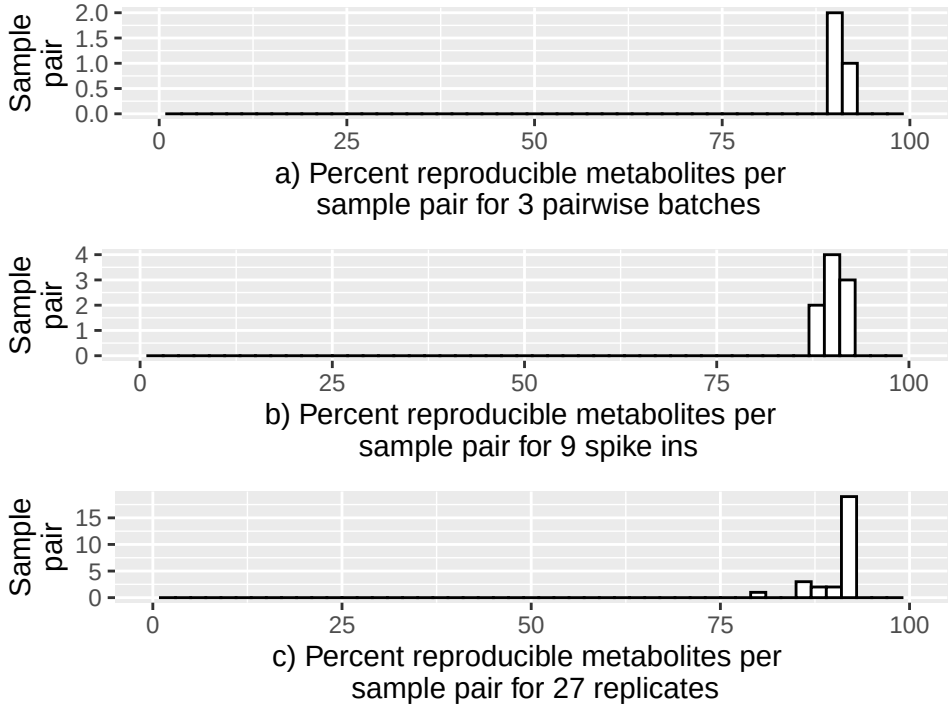

Figure 9: Reproducibility of metabolites per sample pair.

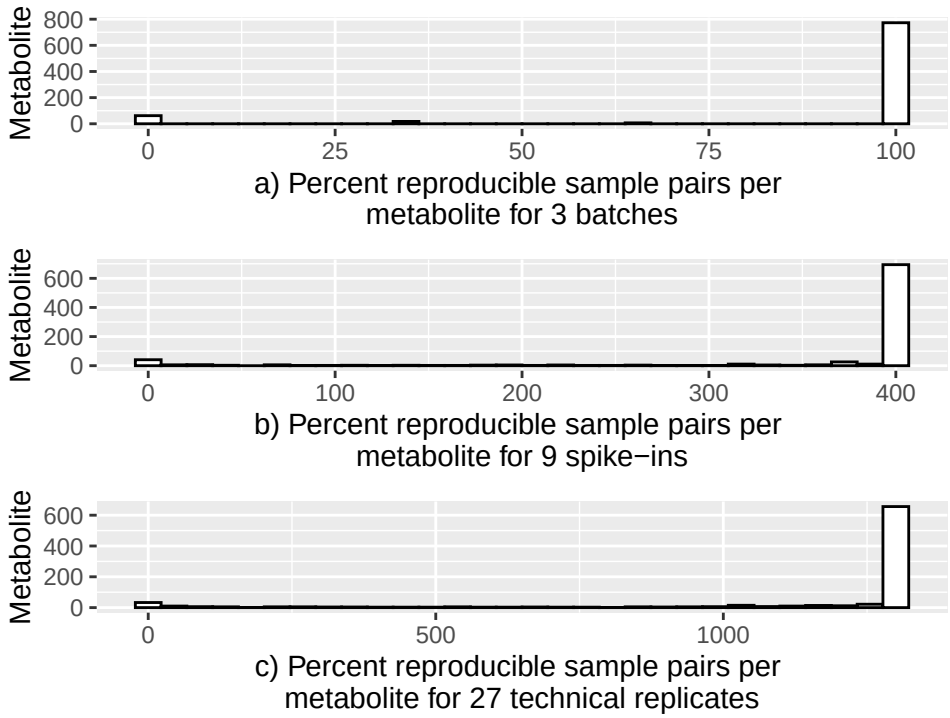

Figure 10: Reproducibility of sample pairs per metabolite.

4.2 BioTech and Bio data

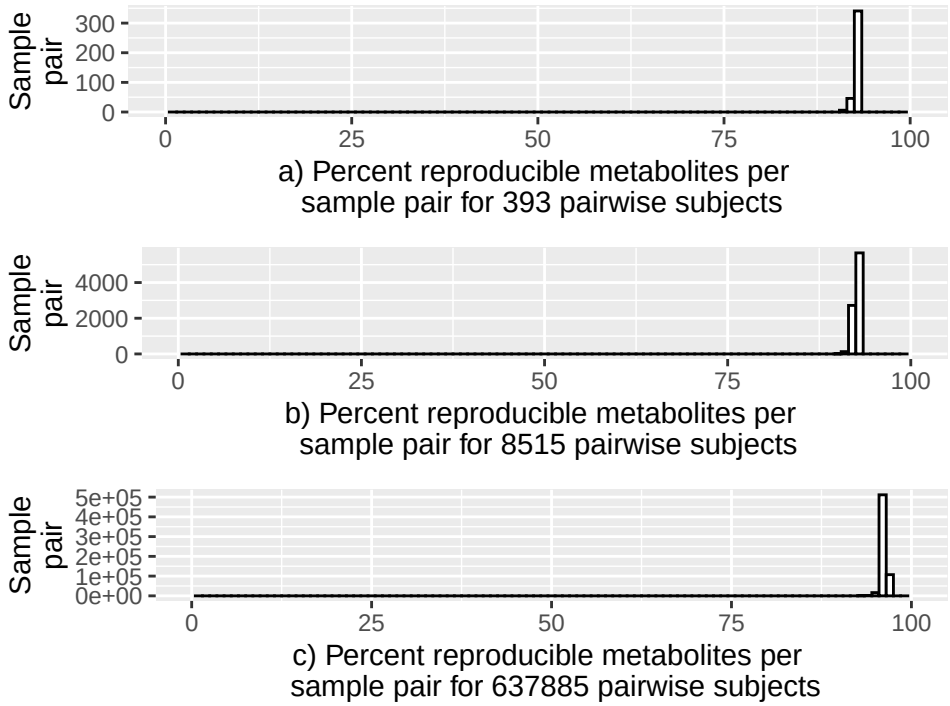

Figure 11: Reproducibility of metabolites per sample pair.

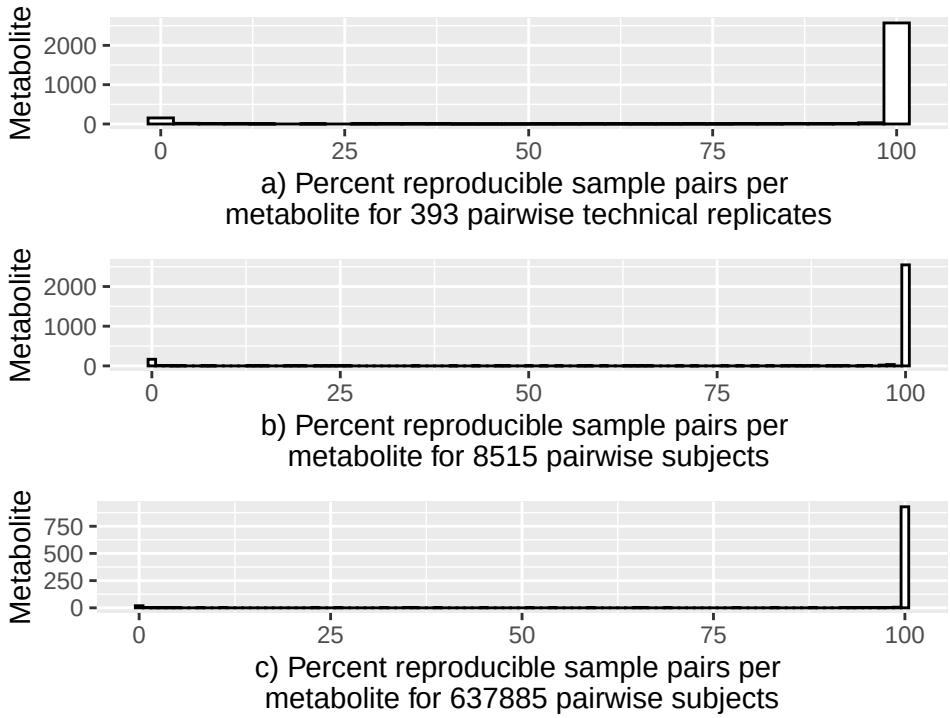

Figure 12: Reproducibility of sample pairs per metabolite.

Data set		Percentage of reproducible metabolites per sample pair			Percentage of reproducible sample pairs per metabolite		
		$> 70\%$	$> 80\%$	$> 90\%$	$> 70\%$	$> 80\%$	$> 90\%$
Tech	Operator layer	100.00	100.00	33.33	89.88	89.88	89.88
	Spike-in layer	100.00	100.00	55.56	92.67	92.67	92.56
	Replicate layer	100.00	100.00	77.78	95.00	94.88	94.65
BioTech	biological layer	100.00	100.00	99.99	92.31	92.10	91.54
	technical layer	100.00	100.00	100.00	92.34	92.03	91.40
Bio	biological layer	100.00	100.00	99.98	95.68	95.58	95.38

Table 3: Summary of reproducible metabolites per sample pair and reproducible sample pairs per metabolite for Tech, BioTech and Bio data sets. The columns of proportion of reproducible metabolites per sample pair $> 70/80/90\%$ indicate that proportion of sample pairs of replicate experiments with reproducible metabolites per sample pair greater than $> 70/80/90\%$. The columns of proportion of reproducible sample pairs per metabolite $> 70/80/90\%$ indicate that proportion of metabolites with reproducible sample pairs per metabolite greater than $> 70/80/90\%$.

5 Imputation- BPCA and Normalization- median

5.1 Tech data

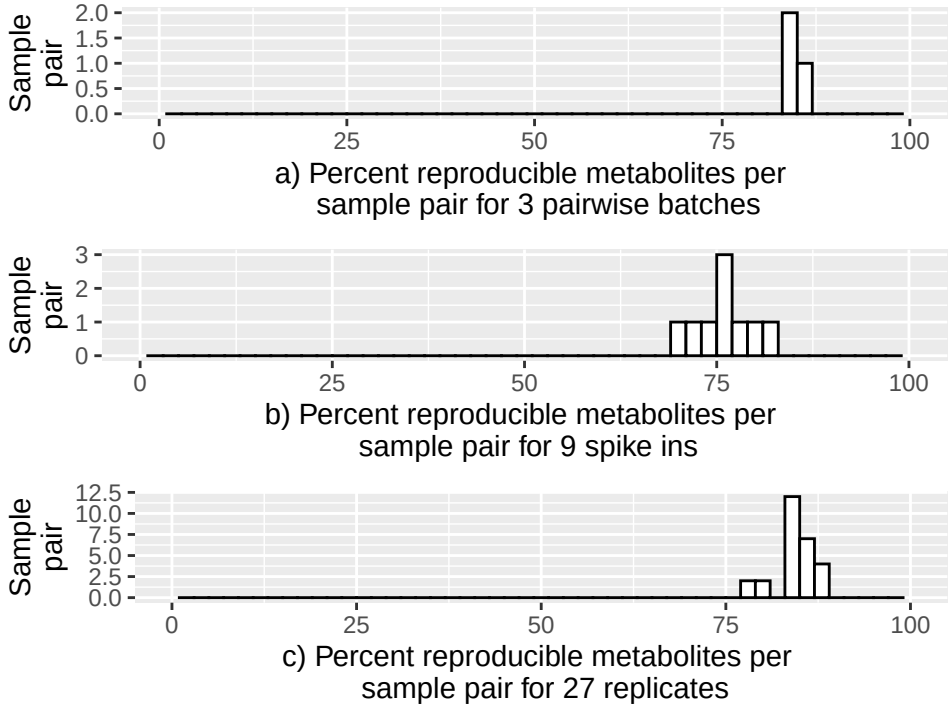

Figure 13: Reproducibility of metabolites per sample pair.

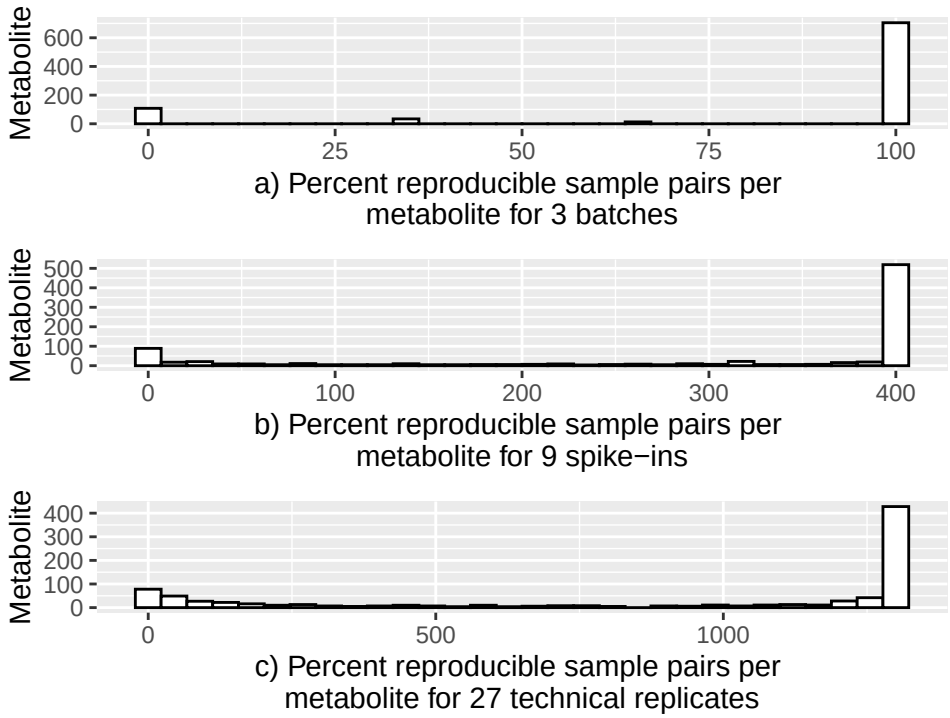

Figure 14: Reproducibility of sample pairs per metabolite.

5.2 BioTech and Bio data

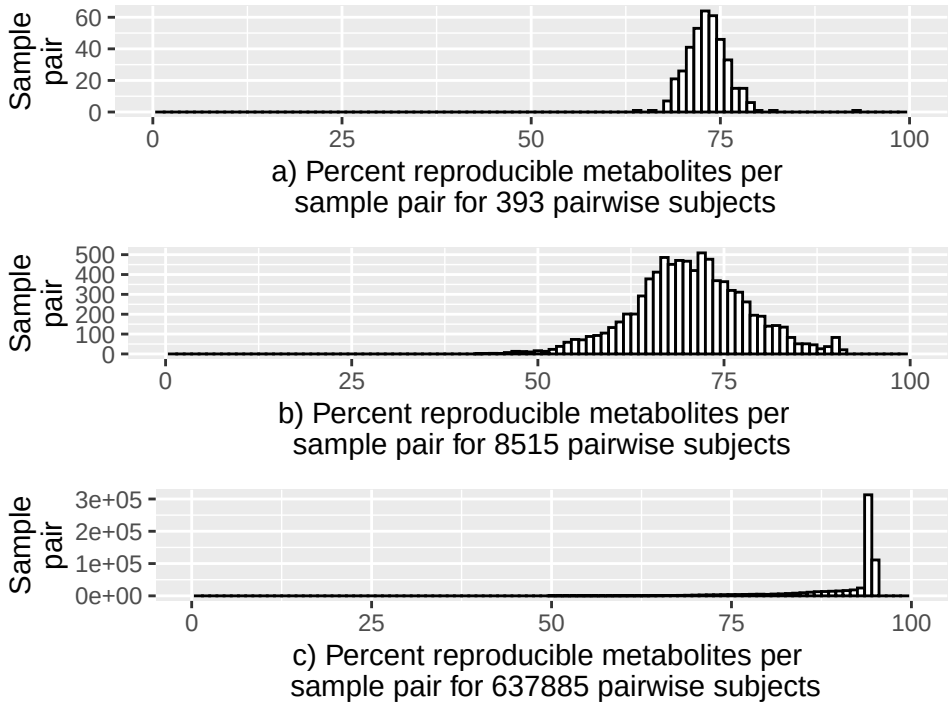

Figure 15: Reproducibility of metabolites per sample pair.

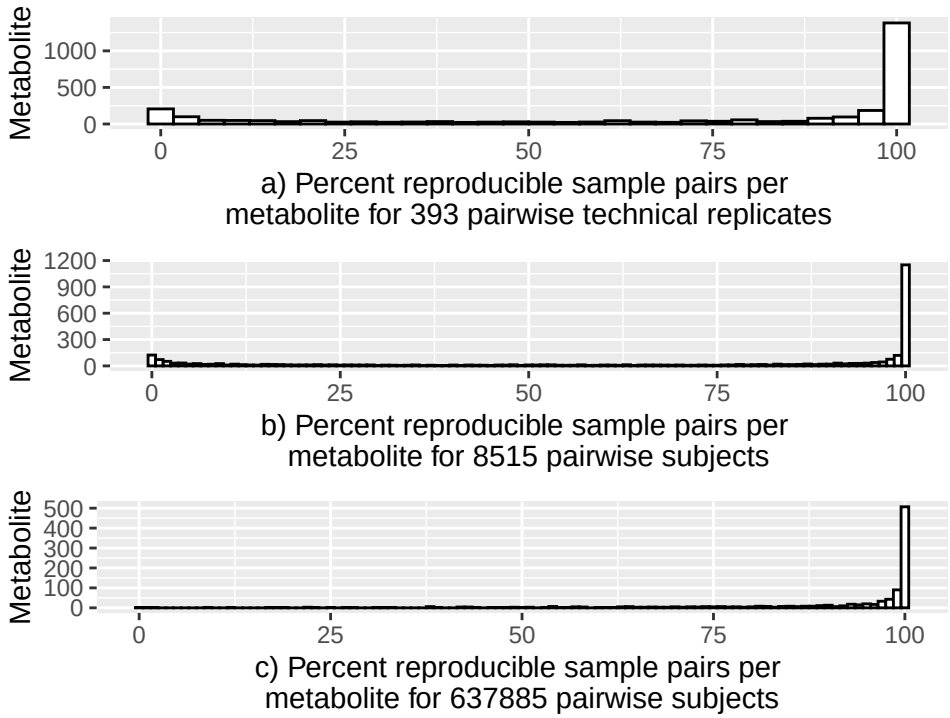

Figure 16: Reproducibility of sample pairs per metabolite.

Data set		Percentage of reproducible metabolites per sample pair			Percentage of reproducible sample pairs per metabolite		
		$> 70\%$	$> 80\%$	$> 90\%$	$> 70\%$	$> 80\%$	$> 90\%$
Tech	Operator layer	100.00	100.00	0.00	81.98	81.98	81.98
	Spike-in layer	88.89	22.22	0.00	82.44	81.74	81.16
	Replicate layer	100.00	92.59	0.00	85.23	84.30	83.02
BioTech	biological layer	52.27	10.68	0.75	64.69	61.22	55.56
	technical layer	90.08	0.76	0.25	68.29	63.78	59.09
Bio	biological layer	98.52	92.40	76.95	88.84	84.32	76.78

Table 4: Summary of reproducible metabolites per sample pair and reproducible sample pairs per metabolite for Tech, BioTech and Bio data sets. The columns of proportion of reproducible metabolites per sample pair $> 70/80/90\%$ indicate that proportion of sample pairs of replicate experiments with reproducible metabolites per sample pair greater than $> 70/80/90\%$. The columns of proportion of reproducible sample pairs per metabolite $> 70/80/90\%$ indicate that proportion of metabolites with reproducible sample pairs per metabolite greater than $> 70/80/90\%$.

6 Imputation- BPCA and Normalization- quantile

6.1 Tech data

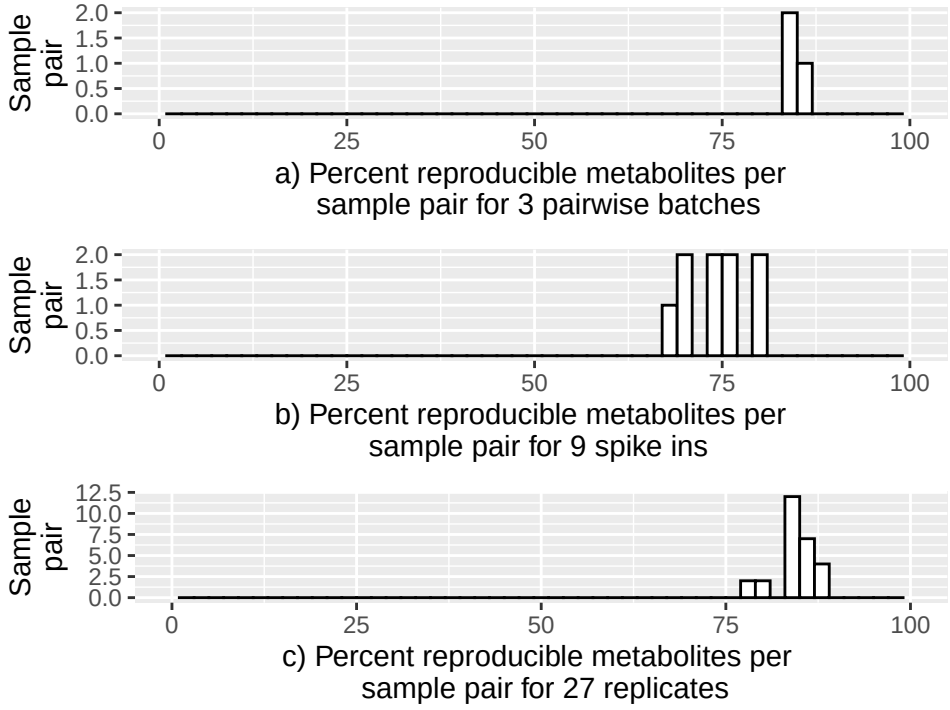

Figure 17: Reproducibility of metabolites per sample pair.

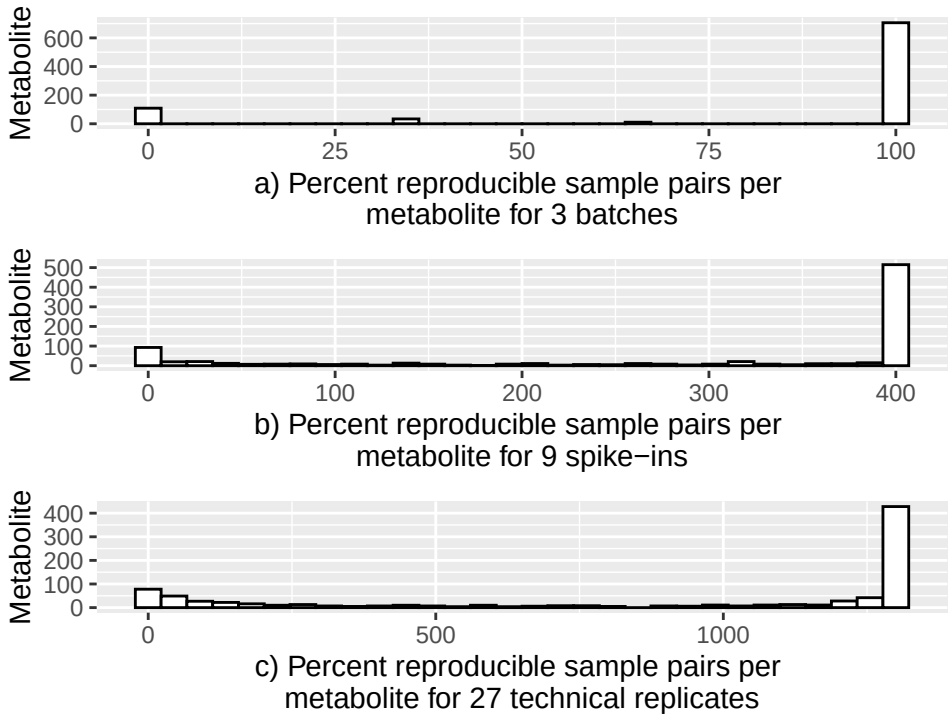

Figure 18: Reproducibility of sample pairs per metabolite.

6.2 BioTech and Bio data

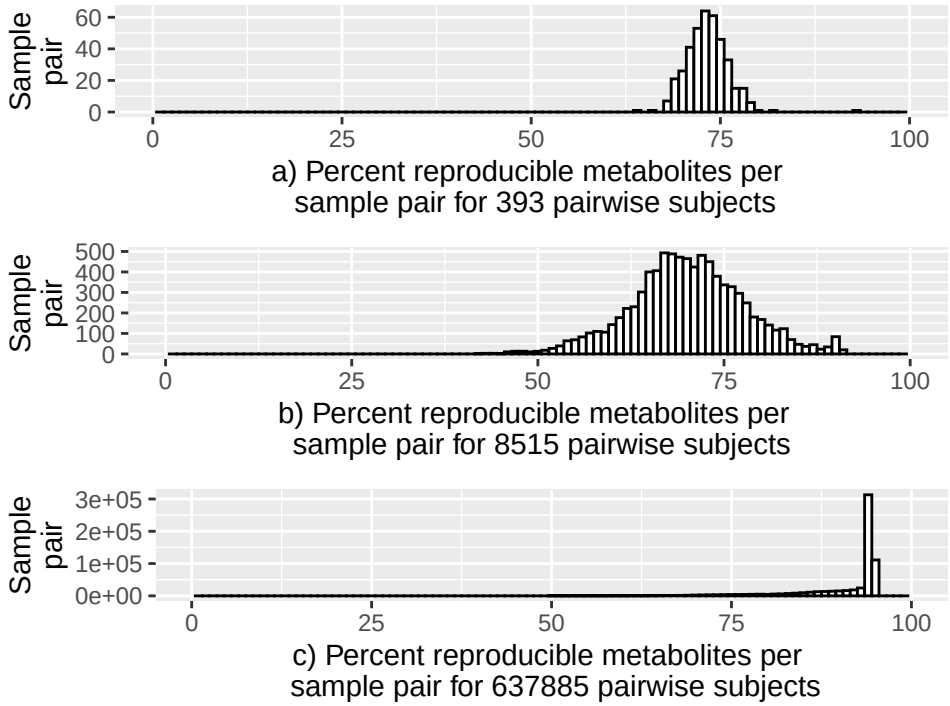

Figure 19: Reproducibility of metabolites per sample pair.

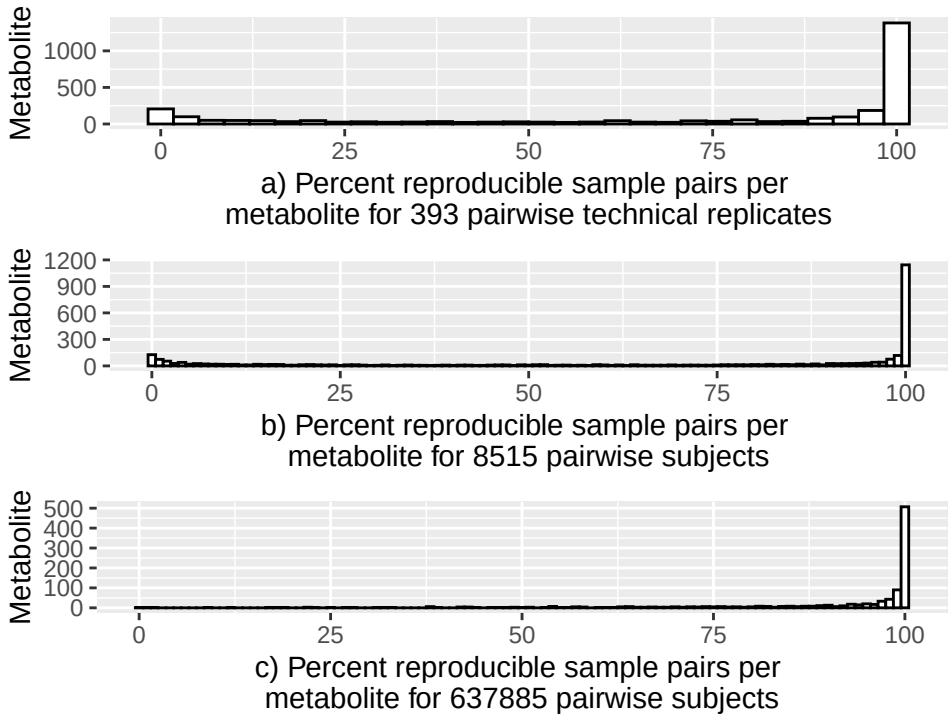

Figure 20: Reproducibility of sample pairs per metabolite.

Data set		Percentage of reproducible metabolites per sample pair			Percentage of reproducible sample pairs per metabolite		
		$> 70\%$	$> 80\%$	$> 90\%$	$> 70\%$	$> 80\%$	$> 90\%$
Tech	Operator layer	100.00	100.00	0.00	82.09	82.09	82.09
	Spike-in layer	88.89	11.11	0.00	81.28	81.05	80.23
	Replicate layer	100.00	92.59	0.00	85.23	84.30	83.02
BioTech	biological layer	49.85	9.59	0.73	64.44	60.80	55.07
	technical layer	90.08	0.76	0.25	68.29	63.78	59.09
Bio	biological layer	98.52	92.40	76.95	88.84	84.32	76.78

Table 5: Summary of reproducible metabolites per sample pair and reproducible sample pairs per metabolite for Tech, BioTech and Bio data sets. The columns of proportion of reproducible metabolites per sample pair $> 70/80/90\%$ indicate that proportion of sample pairs of replicate experiments with reproducible metabolites per sample pair greater than $> 70/80/90\%$. The columns of proportion of reproducible sample pairs per metabolite $> 70/80/90\%$ indicate that proportion of metabolites with reproducible sample pairs per metabolite greater than $> 70/80/90\%$.

7 Imputation- BPCA and Normalization- RUV

7.1 Tech data

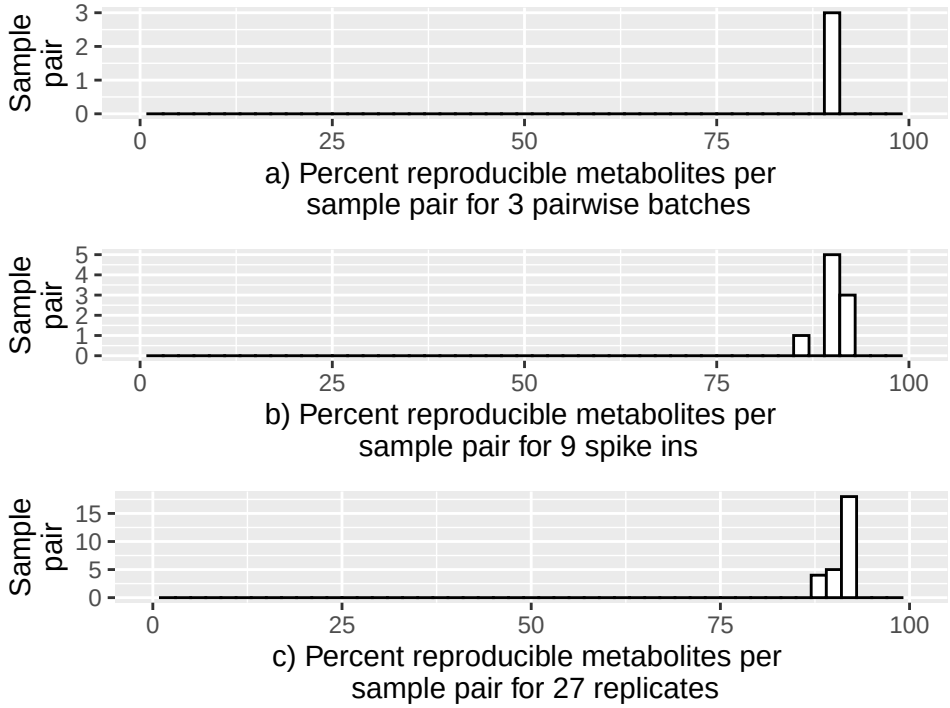

Figure 21: Reproducibility of metabolites per sample pair.

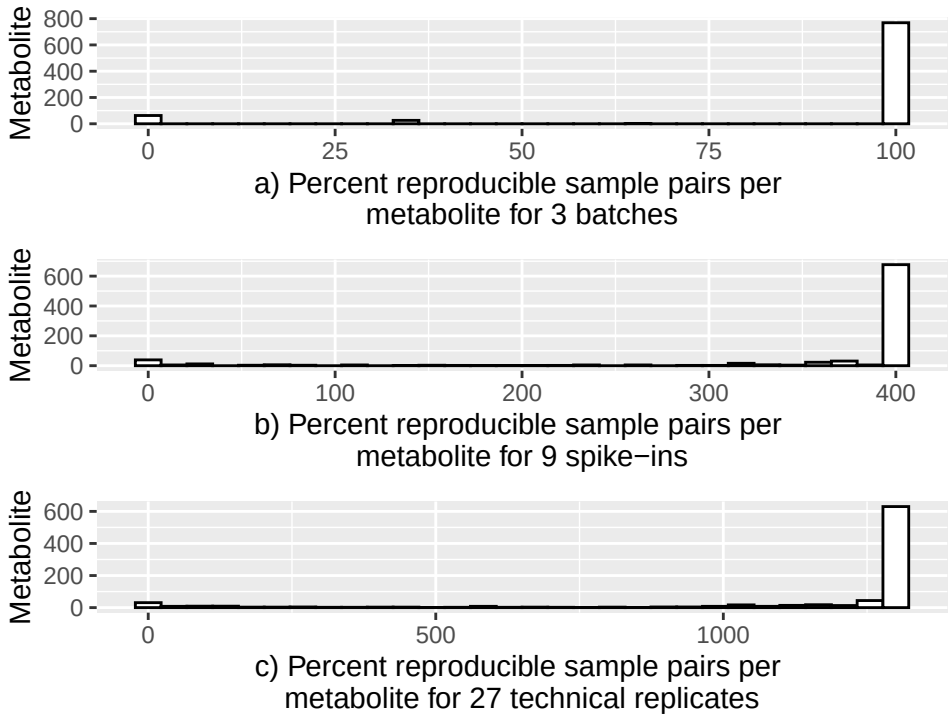

Figure 22: Reproducibility of sample pairs per metabolite.

7.2 BioTech and Bio data

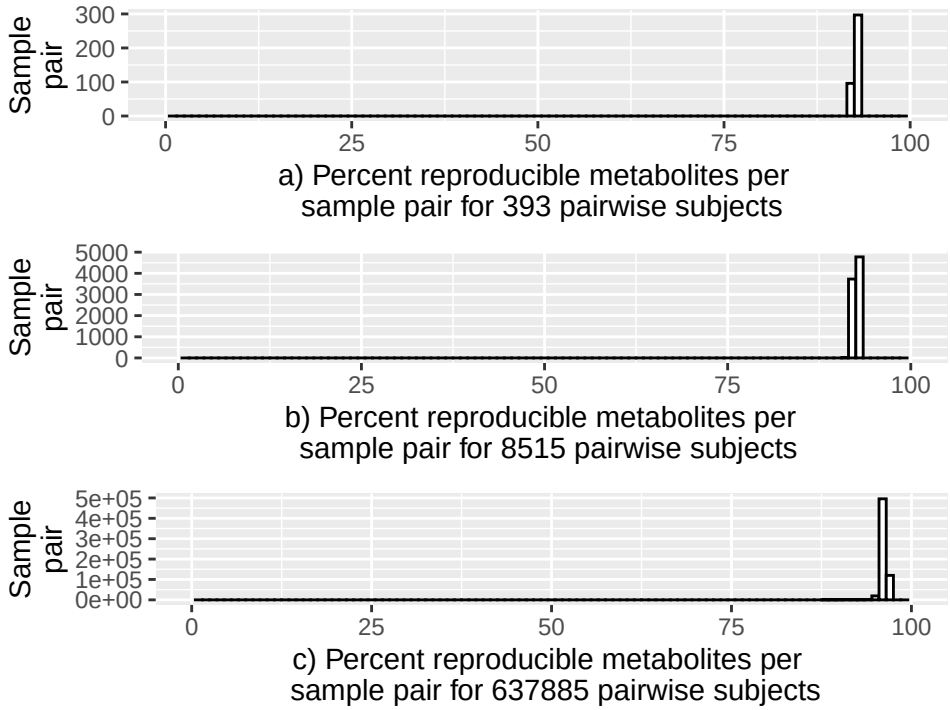

Figure 23: Reproducibility of metabolites per sample pair.

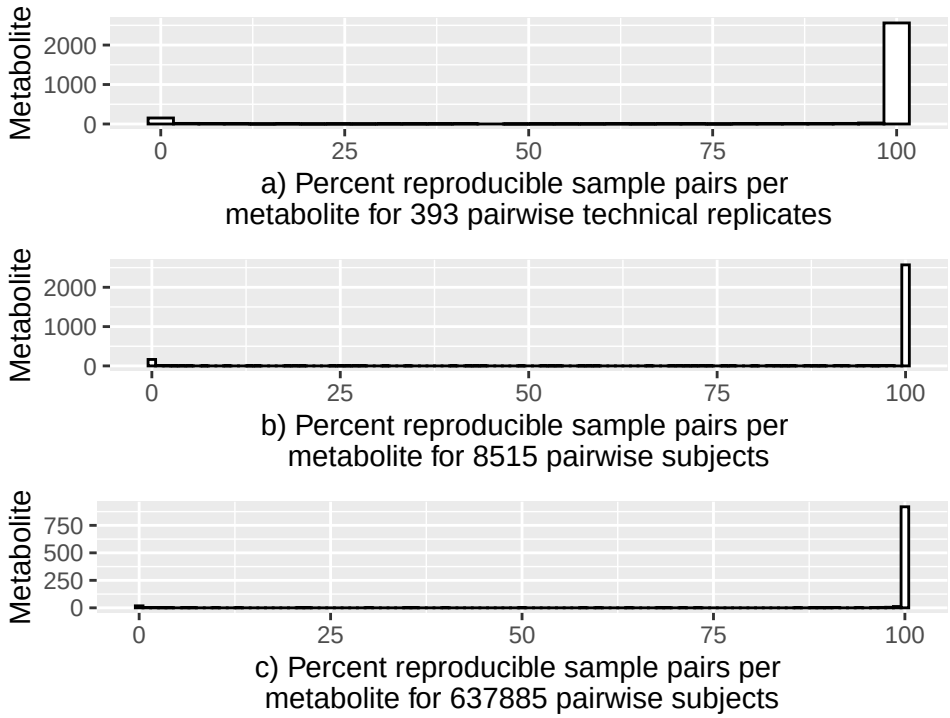

Figure 24: Reproducibility of sample pairs per metabolite.

Data set		Percentage of reproducible metabolites per sample pair			Percentage of reproducible sample pairs per metabolite		
		$> 70\%$	$> 80\%$	$> 90\%$	$> 70\%$	$> 80\%$	$> 90\%$
Tech	Operator layer	100.00	100.00	66.67	89.42	89.42	89.42
	Spike-in layer	100.00	100.00	77.78	92.44	92.21	92.09
	Replicate layer	100.00	100.00	85.19	95.47	95.35	94.77
BioTech	biological layer	100.00	100.00	100.00	92.24	91.85	91.29
	technical layer	100.00	100.00	100.00	92.20	91.78	90.98
Bio	biological layer	100.00	100.00	99.95	95.78	95.58	95.28

Table 6: Summary of reproducible metabolites per sample pair and reproducible sample pairs per metabolite for Tech, BioTech and Bio data sets. The columns of proportion of reproducible metabolites per sample pair $> 70/80/90\%$ indicate that proportion of sample pairs of replicate experiments with reproducible metabolites per sample pair greater than $> 70/80/90\%$. The columns of proportion of reproducible sample pairs per metabolite $> 70/80/90\%$ indicate that proportion of metabolites with reproducible sample pairs per metabolite greater than $> 70/80/90\%$.