BS6213

Batch effects in omics data

Professor Wong Limsoon

6 February 2023

The session looks at a major issue that underlies many omics datasets, viz. batch effects. Batch effects are technical biases that may confound analysis of omics data. They are very complex and effective mitigation is highly context dependent. Do they affect identification of discriminating/causal factors when we analyze patient datasets? Do prediction models (constructed on training datasets) work well on future patients? How do you mitigate batch effects?

Session Plan

Part I, What batch effects are and how they affect biomedical data analysis and model building.

Suggested readings:

- Leek et al., "Tackling the widespread and critical impact of batch effects in high-throughput data", *Nat. Rev. Genet.*, 11(10):733-739, 2010

Part II, How batch effects can be measured. How do you know they are big enough to worry over?

Suggested readings:

- kBET (Buttner et al., "A test metric for assessing single-cell RNA-seq batch corrections", *Nat. Methods*, 16:43-49, 2019)
- PCA side-by-side boxplot (Goh & Wong, "Protein complex-based analysis is resistant to the obfuscating consequences of batch effects – a case study in clinical proteomics", *BMC Genomics*, 18:142, 2017)

Part III, Normalization methods and batch effect-correction methods. What are these and what are their important differences?

Suggested readings:

- Common normalization methods such as linear scaling, quantile normalization, z-score transformation, and specialized methods such as GFS (Belorkar & Wong, "GFS: Fuzzy preprocessing for effective gene expression analysis", *BMC Bioinformatics*, 17(Suppl 17):540, 2016)

- Some popular batch effect-correction methods are ComBat (Johnson et al., “Adjusting batch effects in microarray expression data using empirical Bayes methods”, *Biostatistics*, 8:118-127, 2007), Harman (Otyam et al., “Risk-conscious correction of batch effects: Maximising information extraction from high-throughput genomics datasets”, *BMC Bioinformatics*, 17:332, 2016), SVA (Leek & Storey, “Capturing heterogeneity in gene expression studies by surrogate variable analysis”, *PLoS Genet*, 3:1724-1735, 2007), and Batch mean centering (Sim et al., “The removal of multiplicative, systematic bias allows integration of breast cancer gene expression datasets – improving meta-analysis and prediction of prognosis”, *BMC Med. Genomics*, 1:42, 2008)

Part IV, How should a normalization method be applied when there are multiple classes and batches?

Suggested readings:

- Zhao et al., “How to do quantile normalization correctly for gene expression data analysis”, *Scientific Reports*, 10:15534, 2020

Part V, How do normalization methods interact with batch effects and batch effect-correction methods

Suggested readings:

- Zhou et al., “Examining the practical limits of batch effect-correction algorithms: When should you care about batch effects?”, *J Genet. Genomics*, 46:433-443, 2019.

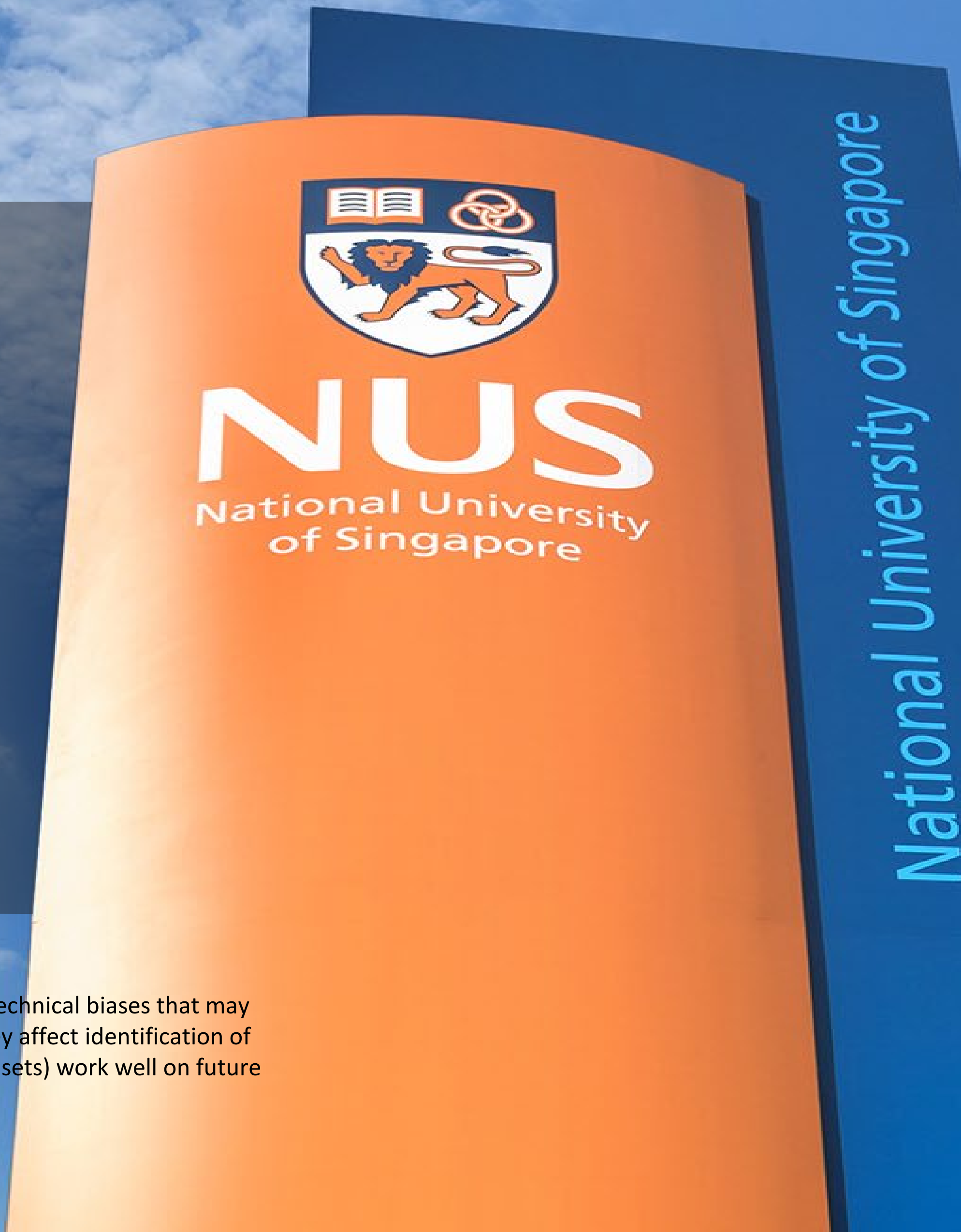
Part VI, If a dataset has lots of missing values and also batch effects, what happens and what can/should you do?

Suggested readings:

- Some missing value-imputation methods (imputation by global mean, same-batch mean, nearest neighbours, etc.)
- Voss et al., “HarmonizR enables data harmonization across independent proteomic datasets with appropriate handling of missing values”, *Nat. Comm.,* 13: 3523, 2022
- Sun & Goh, “Why batch sensitization is important for missing value imputation”, https://doi.org/10.21203/rs.3.rs-1328989/v1

Batch effects in omics data

Wong Limsoon


NUS
National University of Singapore
National University of Singapore

Outline: The session looks at a major issue that underlies many omics datasets, viz. batch effects. Batch effects are technical biases that may confound analysis of omics data. They are very complex and effective mitigation is highly context dependent. Do they affect identification of discriminating/causal factors when we analyze patient datasets? Do prediction models (constructed on training datasets) work well on future patients? How do you mitigate batch effects?

What batch effects are

Batch effects

Batch effects

Unwanted non-biological variations due to processing time, reagent batch, handlers, etc.

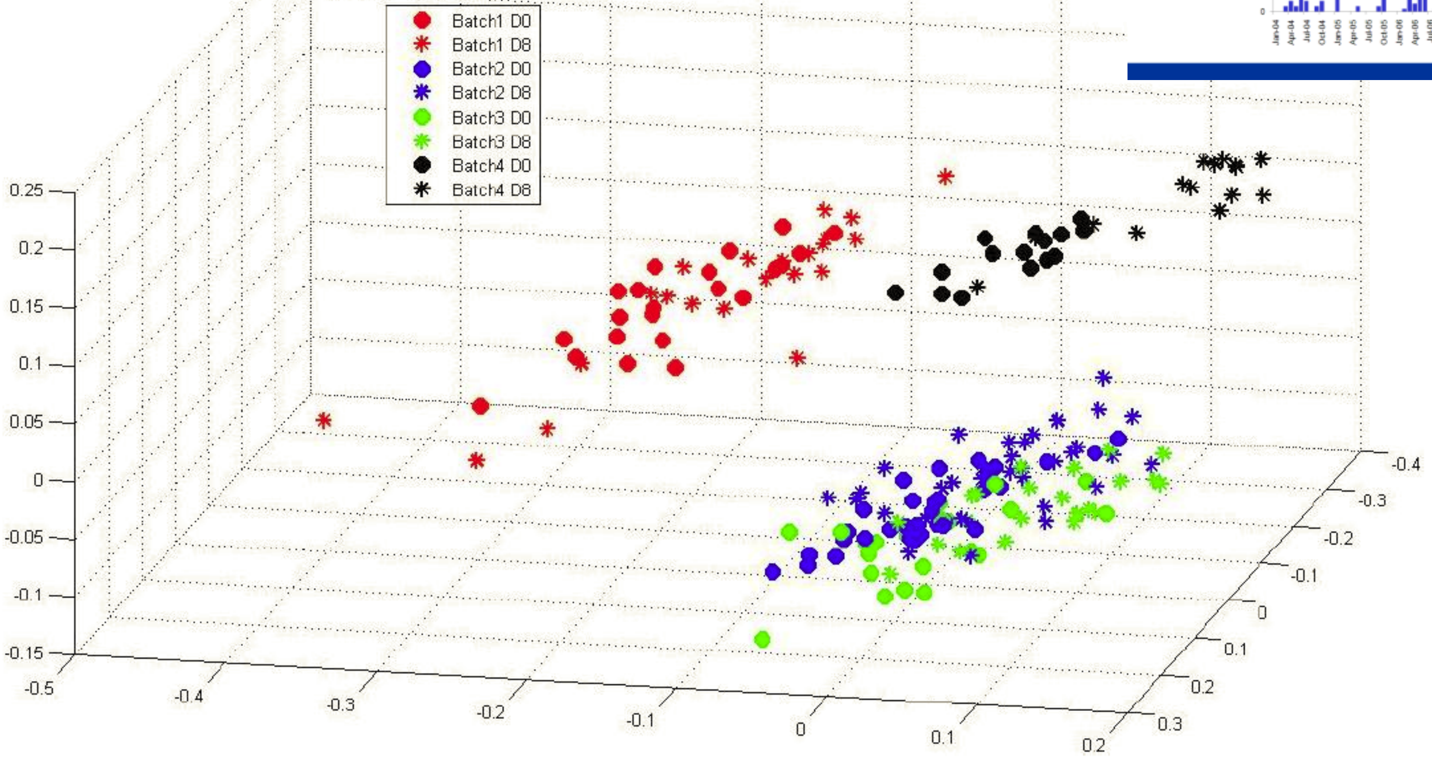
Batch-class imbalance

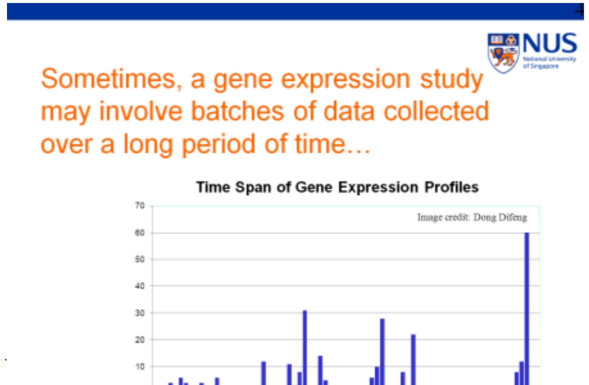
One class forms a large fraction of a batch and another class forms a large fraction of another batch

In this situation, batch effects tend to be badly confounded with biological effects

Childhood leukemia patients

Samples from diff batches are grouped together, regardless of subtypes and treatment response





Sometimes, a gene expression study may involve batches of data collected over a long period of time…

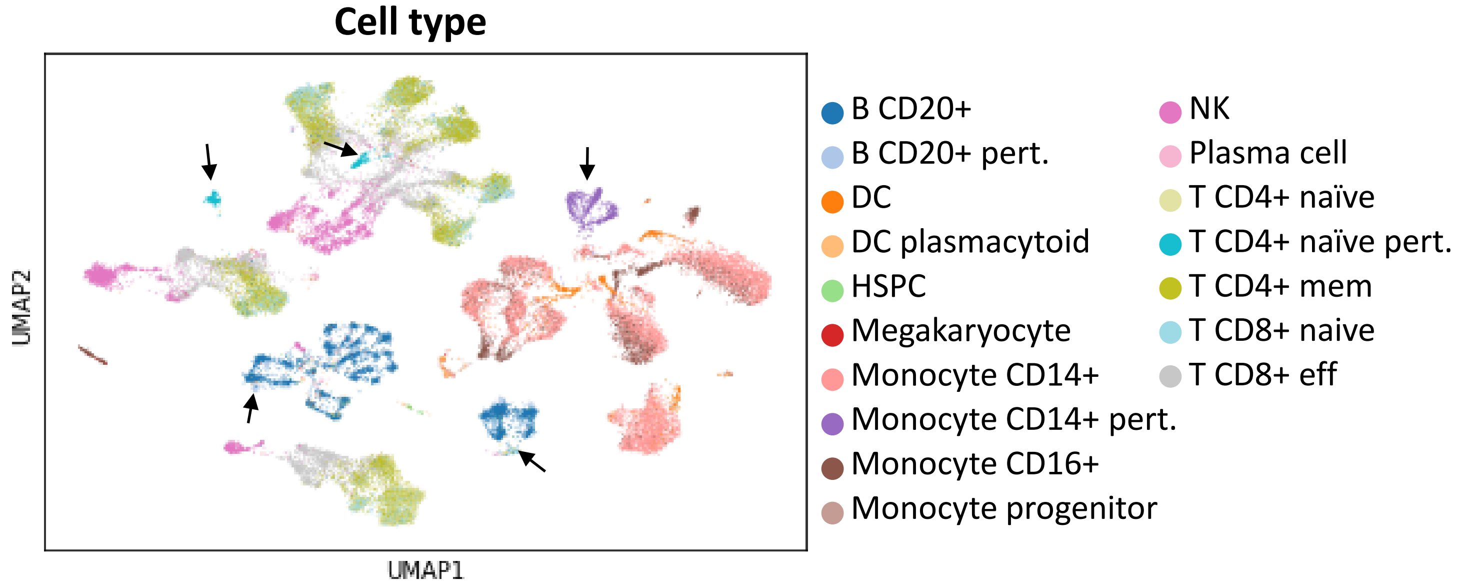
Time Span of Gene Expression Profiles

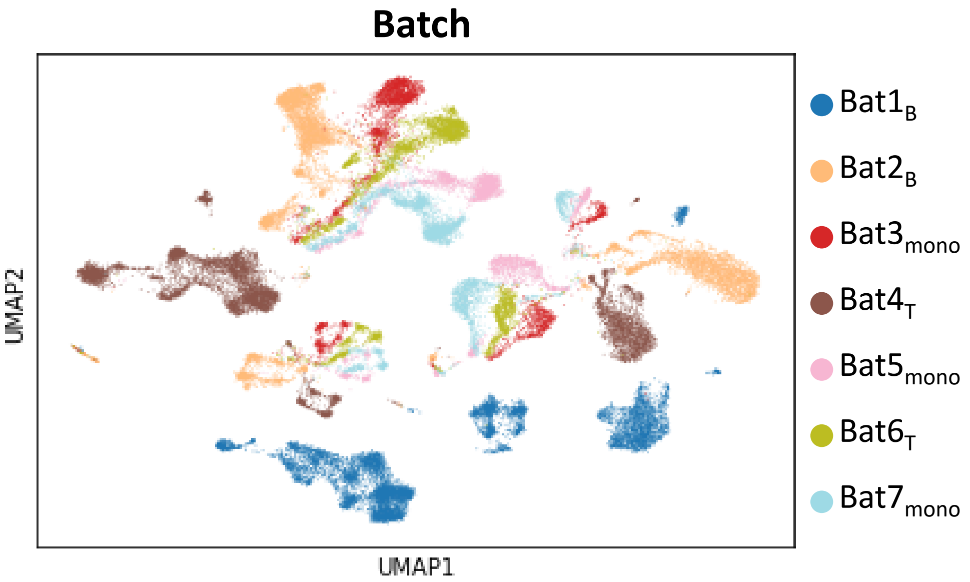
Image credit: Dong Difeng

Copyright 2011 © Limsoon Wong

Image credit: Difeng Dong's PhD dissertation, 2011

Peripheral blood mononuclear cells (PBMC)





Yong et al., *Blood*, 138(S1):2954, 2021

Exercise

Do batch effects affect data analysis and model building?

In what ways?

Intentionally left blank

Exercise

What makes batch-label randomization a valid control?

Intentionally left blank

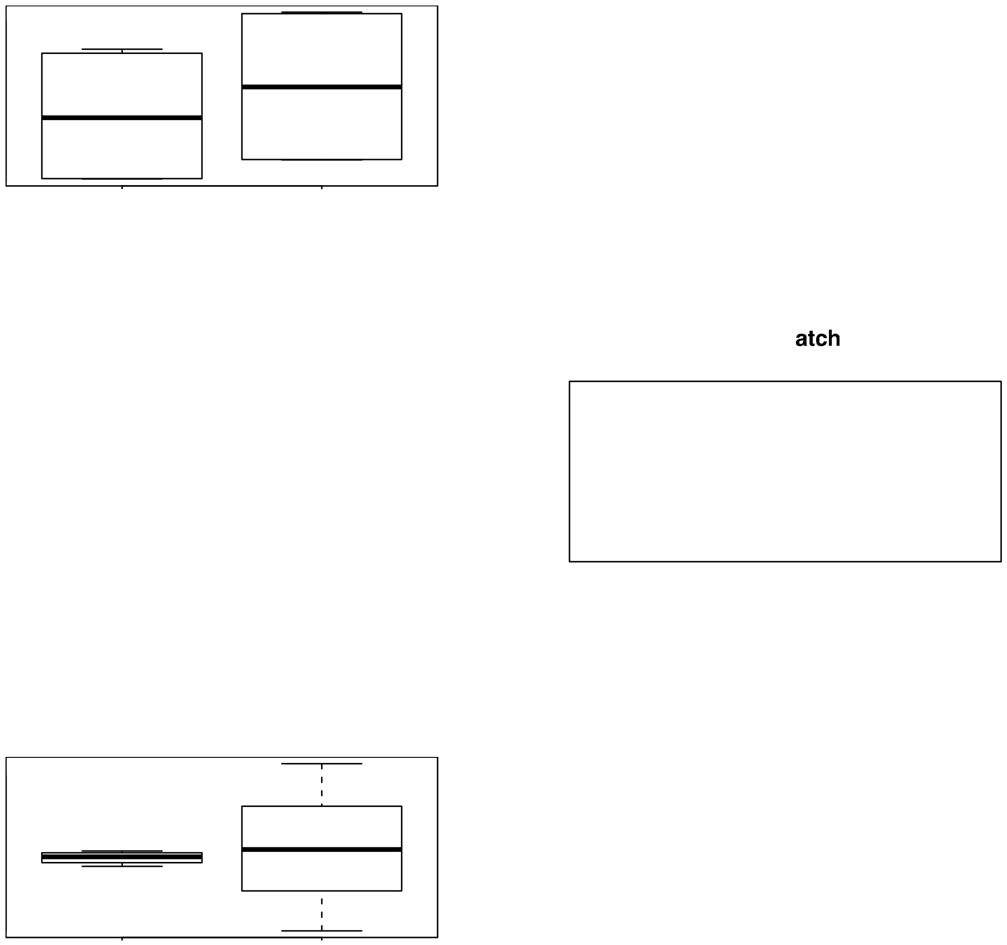
How batch effects are “measured”

Paired boxplots of PCs

Goh & Wong, *BMC Genomics*, 18:142, 2017

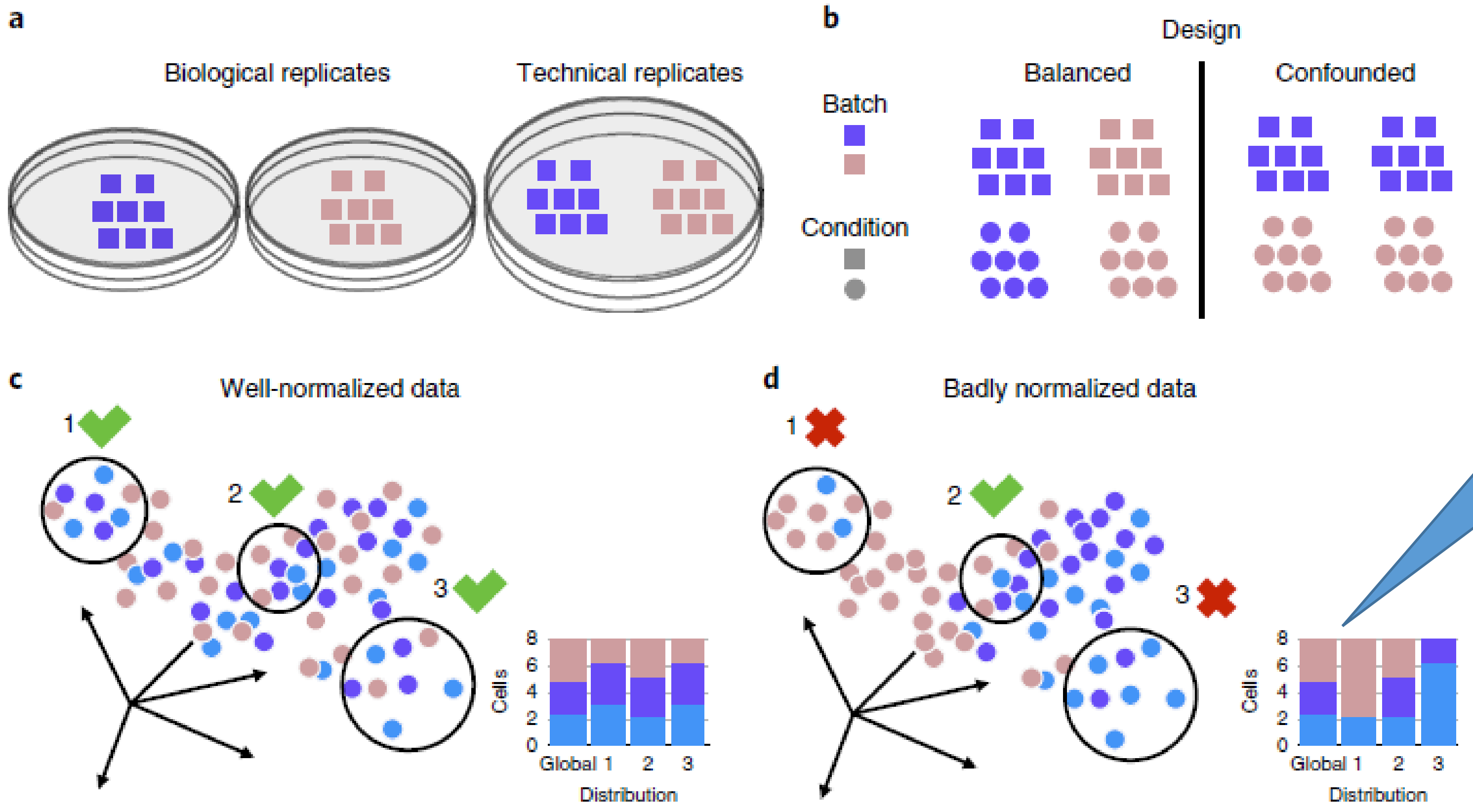
PCA scatter plot is often used for visualizing batch effects

But it is easier to see which PC is enriched in batch effects by showing, side by side, the distribution of values of each PC stratified by class and by batch variables



kBET

Buttner et al., *Nature Methods*, 16:43-49, 2019



$\chi 2$ test the local batch distribution against the global batch distribution

For high-dimensional data, the authors recommend to do PCA, retain the top 50 PCs, then run kBET on the reduced data

Exercise

What is good/bad about paired boxplots of PCs?

What is good/bad about kBET?

E.g., what if class or batch proportions are imbalanced? What if some classes appear only in some batches?

Suggest how to improve either of the above for quantifying batch effects, or suggest a totally different approach

Normalization & batch-effect correction

Normalization vs batch-effect correction

Normalization

Put data into the same scale

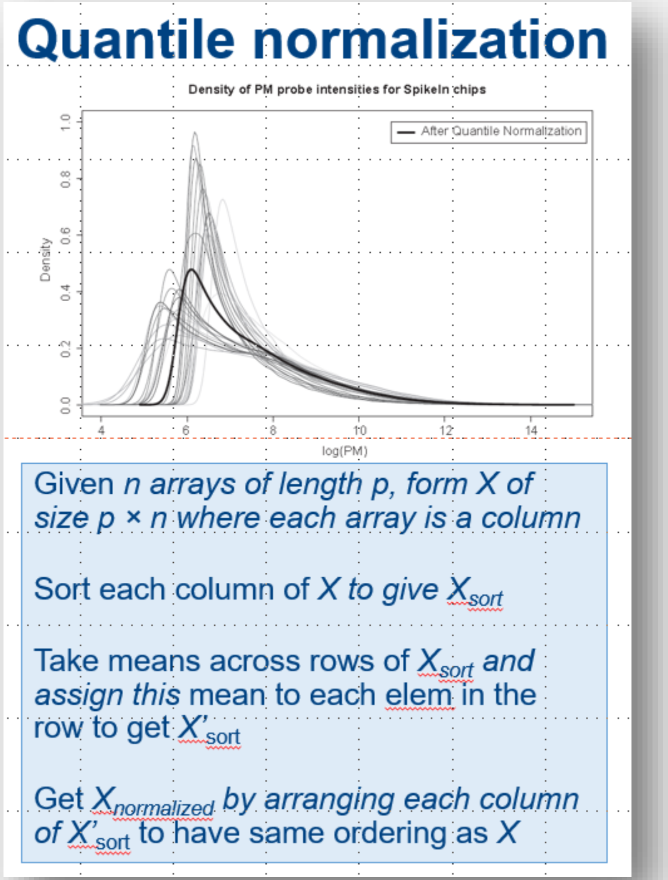
e.g., linear scaling, z-score, quantile normalization, GFS

Batch-effect correction

Remove batch effects

e.g., Combat, Harman, surrogate variable analysis, batch mean centering, GFS

Exercise



Quantile normalization

Density of PM probe intensities for SpikeIn chips
After Quantile Normalization
Density
0.0
0.2
0.4
0.6
0.8
1.0
4
6
8
10
12
14
log(PM)

Given *n arrays of length p, form X of size p × n where each array is a column*

Sort each column of *X to give* X_{sort}

Take means across rows of X_{sort} *and assign this* mean to each elem in the row to get X'_{sort}

Get $X_{normalized}$ *by arranging each column of* X'_{sort} to have same ordering as *X*

Does quantile normalization remove batch effects?

Does it make it easier to identify differentially expressed genes?

Exercise

Look up batch mean centering (BMC)

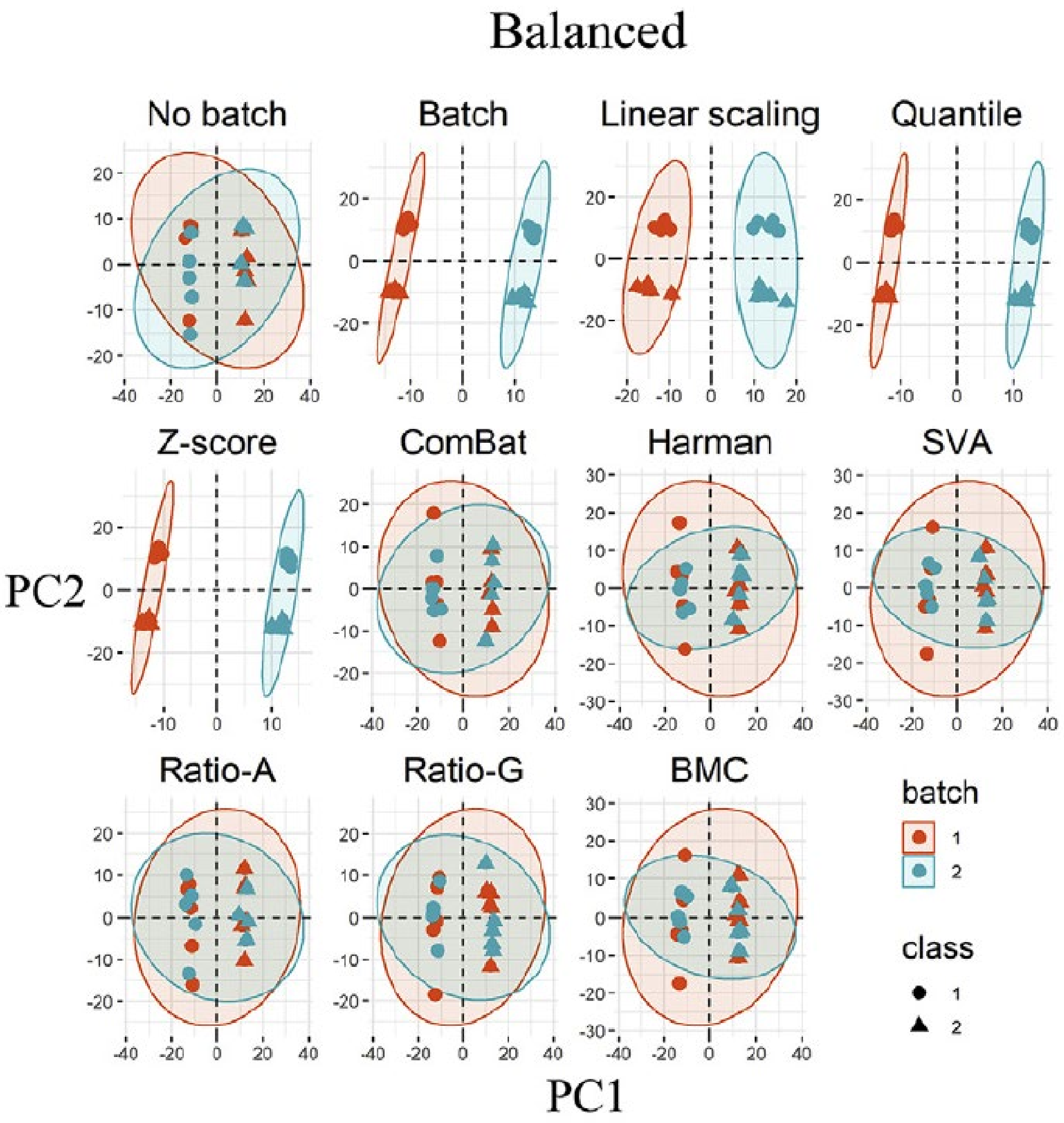
Does it remove additive batch effects well?

Does it remove multiplicative batch effects well?

When class & batch are balanced

Normalization methods (e.g., quantile) do not remove batch effects

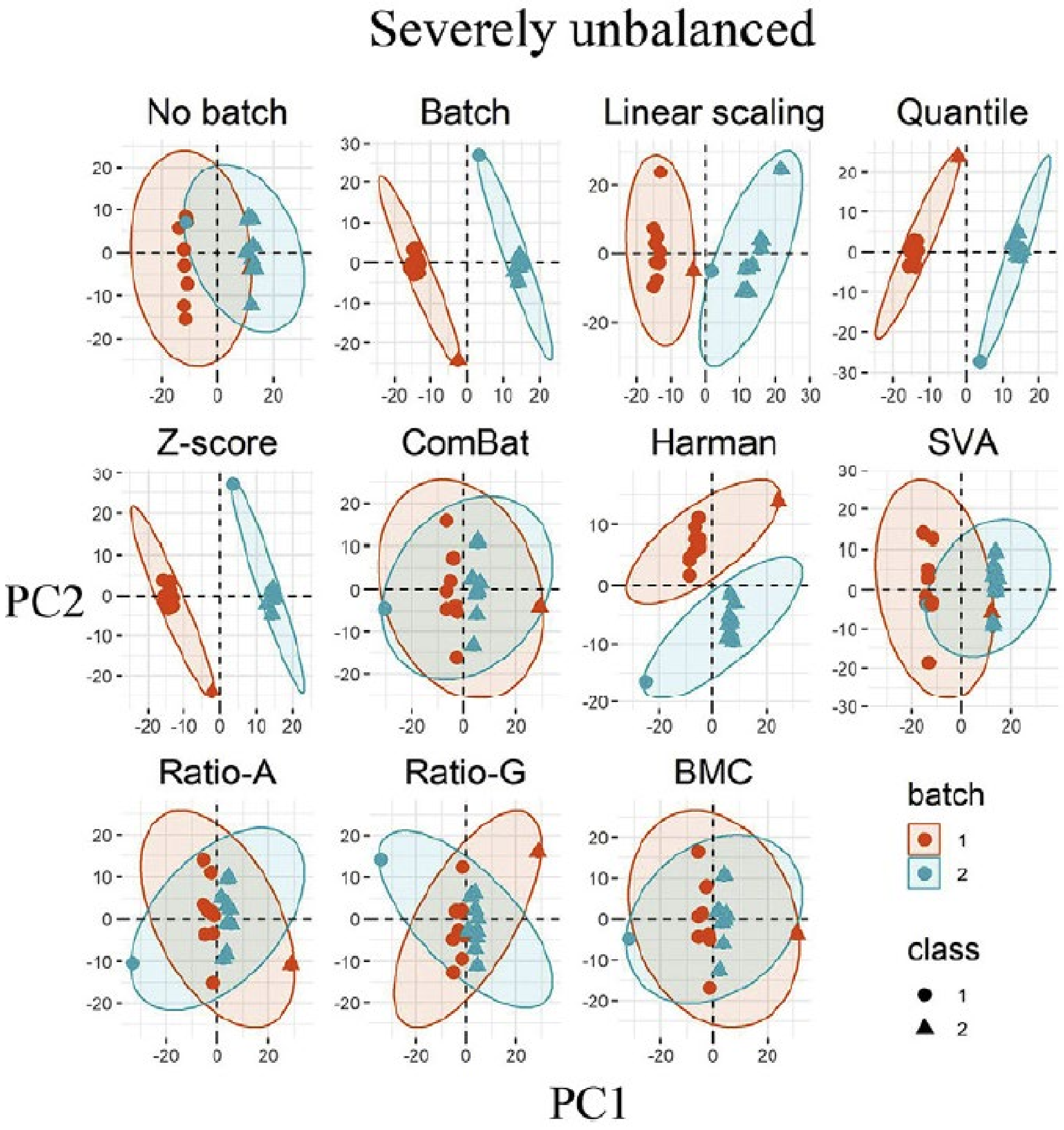
But they still easily separate the classes



Zhou et al., *J Genet & Genom*, 46:433-443, 2019

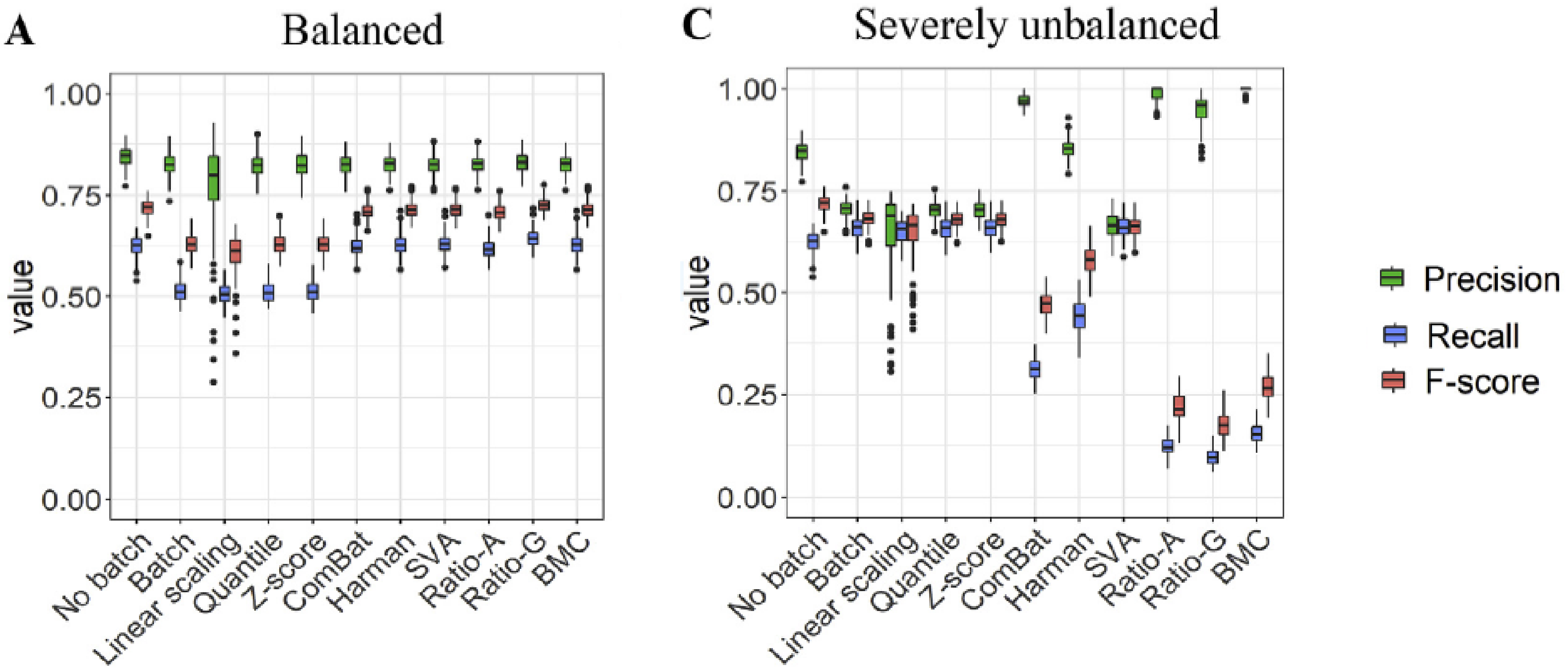
When class & batch are imbalanced, i.e., when one batch is dominated by one class

The situation deteriorates quickly …



Zhou et al., *J Genet & Genom*, 46:433-443, 2019

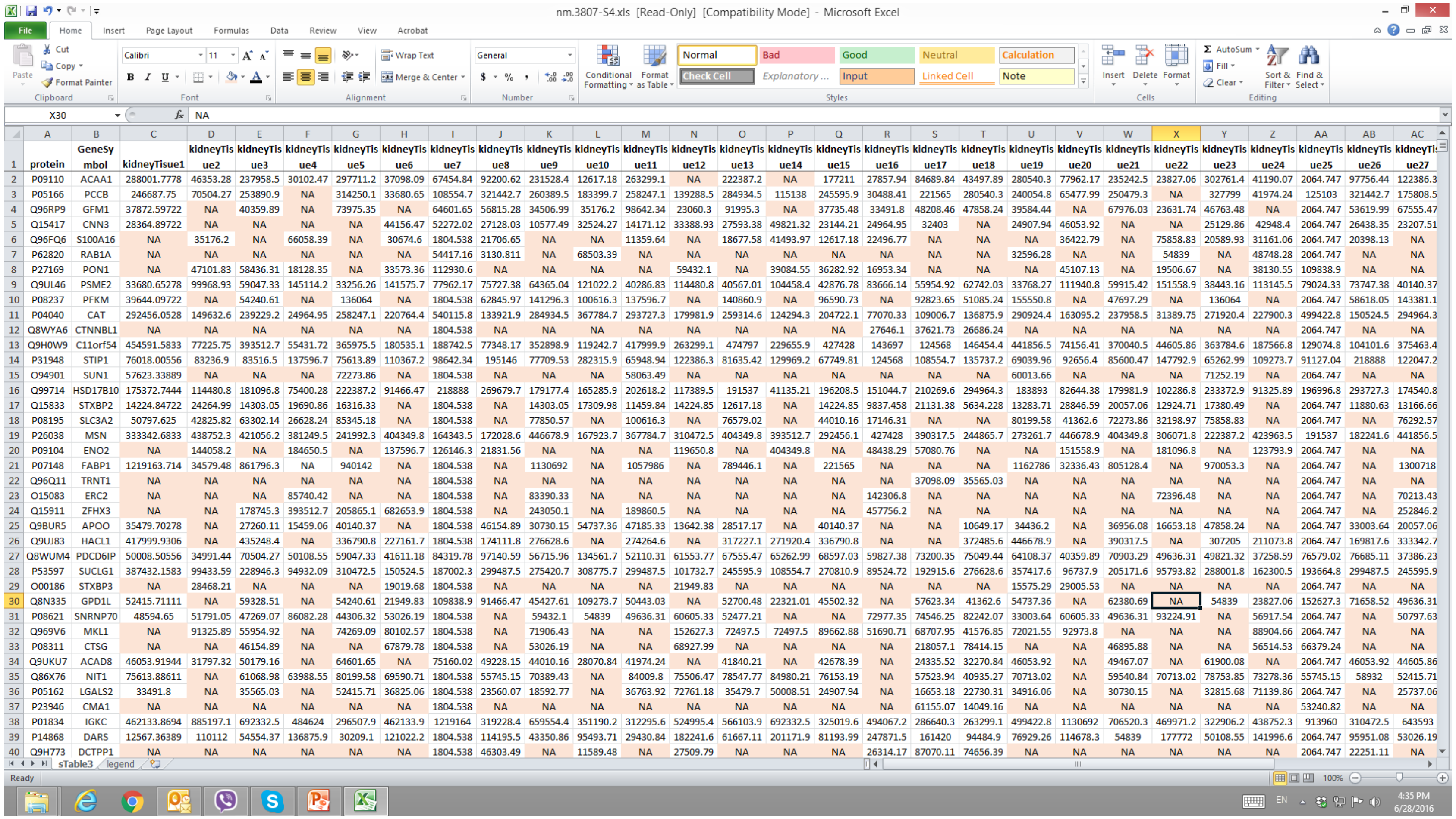
Impact on feature selection



Zhou et al., *J Genet & Genom*, 46:433-443, 2019

Missing values & batch effects

Some omics data have lots of missing values (proteomics MS, scRNA-seq, etc.)



nm.3807-S4.xls [Read-Only] [Compatibility Mode] - Microsoft Excel

X30 NA

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC
1	protein	GeneSy mbol	kidneyTisue1	kidneyTis ue2	kidneyTis ue3	kidneyTis ue4	kidneyTis ue5	kidneyTis ue6	kidneyTis ue7	kidneyTis ue8	kidneyTis ue9	kidneyTis ue10	kidneyTis ue11	kidneyTis ue12	kidneyTis ue13	kidneyTis ue14	kidneyTis ue15	kidneyTis ue16	kidneyTis ue17	kidneyTis ue18	kidneyTis ue19	kidneyTis ue20	kidneyTis ue21	kidneyTis ue22	kidneyTis ue23	kidneyTis ue24	kidneyTis ue25	kidneyTis ue26	kidneyTi ue27
2	P09110	ACAA1	288001.7778	46353.28	237958.5	30102.47	297711.2	37098.09	67454.84	92200.62	231528.4	12617.18	263299.1	NA	222387.2	NA	177211	27857.94	84689.84	43497.89	280540.3	77962.17	235242.5	23827.06	302761.4	41190.07	2064.747	97756.44	122386.3
3	P05166	PCCB	246687.75	70504.27	253890.9	NA	314250.1	33680.65	108554.7	321442.7	260389.5	183399.7	258247.1	139288.5	284934.5	115138	245595.9	30488.41	221565	280540.3	240054.8	65477.99	250479.3	NA	327799	41974.24	125103	321442.7	175808.5
4	Q96RP9	GFM1	37872.59722	NA	40359.89	NA	73975.35	NA	64601.65	56815.28	34506.99	35176.2	98642.34	23060.3	91995.3	NA	37735.48	33491.8	48208.46	47858.24	39584.44	NA	67976.03	23631.74	46763.48	NA	2064.747	53619.99	67555.47
5	Q15417	CNN3	28364.89722	NA	NA	NA	NA	44156.47	52272.02	27128.03	10577.49	32524.27	14171.12	33388.93	27593.38	49821.32	23144.21	24964.95	32403	NA	24907.94	46053.92	NA	NA	25129.86	42948.4	2064.747	26438.35	23207.51
6	Q96FQ6	S100A16	NA	35176.2	NA	66058.39	NA	30674.6	1804.538	21706.65	NA	NA	11359.64	NA	18677.58	41493.97	12617.18	22496.77	NA	NA	NA	36422.79	NA	75858.83	20589.93	31161.06	2064.747	20398.13	NA
7	P62820	RAB1A	NA	NA	NA	NA	NA	NA	54417.16	3130.811	NA	68503.39	NA	NA	NA	NA	NA	NA	NA	NA	32596.28	NA	NA	54839	NA	48748.28	2064.747	NA	NA
8	P27169	PON1	NA	47101.83	58436.31	18128.35	NA	33573.36	112930.6	NA	NA	NA	NA	59432.1	NA	39084.55	36282.92	16953.34	NA	NA	NA	45107.13	NA	19506.67	NA	38130.55	109838.9	NA	NA
9	Q9UL46	PSME2	33680.65278	99968.93	59047.33	145114.2	33256.26	141575.7	77962.17	75727.38	64365.04	121022.2	40286.83	114480.8	40567.01	104458.4	42876.78	83666.14	55954.92	62742.03	33768.27	111940.8	59915.42	151558.9	38443.16	113145.5	79024.33	73747.38	40140.37
10	P08237	PFKM	39644.09722	NA	54240.61	NA	136064	NA	1804.538	62845.97	141296.3	100616.3	137596.7	NA	140860.9	NA	96590.73	NA	92823.65	51085.24	155550.8	NA	47697.29	NA	136064	NA	2064.747	58618.05	143381.1
11	P04040	CAT	292456.0528	149632.6	239229.2	24964.95	258247.1	220764.4	540115.8	133921.9	284934.5	367784.7	293727.3	179981.9	259314.6	124294.3	204722.1	77070.33	109006.7	136875.9	290924.4	163095.2	237958.5	31389.75	271920.4	227900.3	499422.8	150524.5	294964.3
12	Q8WYA6	CTNNBL1	NA	NA	NA	NA	NA	NA	1804.538	NA	NA	NA	NA	NA	NA	NA	NA	27646.1	37621.73	26686.24	NA	NA	NA	NA	NA	NA	2064.747	NA	NA
13	Q9H0W9	C11orf54	454591.5833	77225.75	393512.7	55431.72	365975.5	180535.1	188742.5	77348.17	352898.9	119242.7	417999.9	263299.1	474797	229655.9	427428	143697	124568	146454.4	441856.5	74156.41	370040.5	44605.86	363784.6	187566.8	129074.8	104101.6	375463.4
14	P31948	STIP1	76018.00556	83236.9	83516.5	137596.7	75613.89	110367.2	98642.34	195146	77709.53	282315.9	65948.94	122386.3	81635.42	129969.2	67749.81	124568	108554.7	135737.2	69039.96	92656.4	85600.47	147792.9	65262.99	109273.7	91127.04	218888	122047.2
15	O94901	SUN1	57623.33889	NA	NA	NA	72273.86	NA	1804.538	NA	NA	NA	58063.49	NA	NA	NA	NA	NA	NA	NA	60013.66	NA	NA	NA	71252.19	NA	2064.747	NA	NA
16	Q99714	HSD17B10	175372.7444	114480.8	181096.8	75400.28	222387.2	91466.47	218888	269679.7	179177.4	165285.9	202618.2	117389.5	191537	41135.21	196208.5	151044.7	210269.6	294964.3	183893	82644.38	179981.9	102286.8	233372.9	91325.89	196996.8	293727.3	174540.8
17	Q15833	STXBP2	14224.84722	24264.99	14303.05	19690.86	16316.33	NA	1804.538	NA	14303.05	17309.98	11459.84	14224.85	12617.18	NA	14224.85	9837.458	21131.38	5634.228	13283.71	28846.59	20057.06	12924.71	17380.49	NA	2064.747	11880.63	13166.66
18	P08195	SLC3A2	50797.625	42825.82	63302.14	26628.24	85345.18	NA	1804.538	NA	77850.57	NA	100616.3	NA	76579.02	NA	44010.16	17146.31	NA	NA	80199.58	41362.6	72273.86	32198.97	75858.83	NA	2064.747	NA	76292.57
19	P26038	MSN	333342.6833	438752.3	421056.2	381249.5	241992.3	404349.8	164343.5	172028.6	446678.9	167923.7	367784.7	310472.5	404349.8	393512.7	292456.1	427428	390317.5	244865.7	273261.7	446678.9	404349.8	306071.8	222387.2	423963.5	191537	182241.6	441856.5
20	P09104	ENO2	NA	144058.2	NA	184650.5	NA	137596.7	126146.3	21831.56	NA	NA	NA	119650.8	NA	404349.8	NA	48438.29	57080.76	NA	NA	151558.9	NA	181096.8	NA	123793.9	2064.747	NA	NA
21	P07148	FABP1	1219163.714	34579.48	861796.3	NA	940142	NA	1804.538	NA	1130692	NA	1057986	NA	789446.1	NA	221565	NA	NA	NA	1162786	32336.43	805128.4	NA	970053.3	NA	2064.747	NA	1300718
22	Q96Q11	TRNT1	NA	NA	NA	NA	NA	NA	1804.538	NA	NA	NA	NA	NA	NA	NA	NA	NA	37098.09	35565.03	NA	NA	NA	NA	NA	NA	2064.747	NA	NA
23	O15083	ERC2	NA	NA	NA	85740.42	NA	NA	1804.538	NA	83390.33	NA	NA	NA	NA	NA	NA	142306.8	NA	NA	NA	NA	NA	72396.48	NA	NA	2064.747	NA	70213.43
24	Q15911	ZFHX3	NA	NA	178745.3	393512.7	205865.1	682653.9	1804.538	NA	243050.1	NA	189860.5	NA	NA	NA	NA	457756.2	NA	NA	NA	NA	NA	NA	NA	NA	2064.747	NA	252846.2
25	Q9BUR5	APOO	35479.70278	NA	27260.11	15459.06	40140.37	NA	1804.538	46154.89	30730.15	54737.36	47185.33	13642.38	28517.17	NA	40140.37	NA	NA	10649.17	34436.2	NA	36956.08	16653.18	47858.24	NA	2064.747	33003.64	20057.06
26	Q9UJ83	HACL1	417999.9306	NA	435248.4	NA	336790.8	227161.7	1804.538	174111.8	276628.6	NA	274264.6	NA	317227.1	271920.4	336790.8	NA	NA	372485.6	446678.9	NA	390317.5	NA	307205	211073.8	2064.747	169817.6	333342.7
27	Q8WUM4	PDCD6IP	50008.50556	34991.44	70504.27	50108.55	59047.33	41611.18	84319.78	97140.59	56715.96	134561.7	52110.31	61553.77	67555.47	65262.99	68597.03	59827.38	73200.35	75049.44	64108.37	40359.89	70903.29	49636.31	49821.32	37258.59	76579.02	76685.11	37386.23
28	P53597	SUCLG1	387432.1583	99433.59	228946.3	94932.09	310472.5	150524.5	187002.3	299487.5	275420.7	308775.7	299487.5	101732.7	245595.9	108554.7	270810.9	89524.72	192915.6	276628.6	357417.6	96737.9	205171.6	95793.82	288001.8	162300.5	193664.8	299487.5	245595.9
29	O00186	STXBP3	NA	28468.21	NA	NA	NA	19019.68	1804.538	NA	NA	NA	NA	21949.83	NA	NA	NA	NA	NA	NA	15575.29	29005.53	NA	NA	NA	NA	2064.747	NA	NA
30	Q8N335	GPD1L	52415.71111	NA	59328.51	NA	54240.61	21949.83	109838.9	91466.47	45427.61	109273.7	50443.03	NA	52700.48	22321.01	45502.32	NA	57623.34	41362.6	54737.36	NA	62380.69	NA	54839	23827.06	152627.3	71658.52	49636.31
31	P08621	SNRNP70	48594.65	51791.05	47269.07	86082.28	44306.32	53026.19	1804.538	NA	59432.1	54839	49636.31	60605.33	52477.21	NA	NA	72977.35	74546.25	82242.07	33003.64	60605.33	49636.31	93224.91	NA	56917.54	2064.747	NA	50797.63
32	Q969V6	MKL1	NA	91325.89	55954.92	NA	74269.09	80102.57	1804.538	NA	71906.43	NA	NA	152627.3	72497.5	72497.5	89662.88	51690.71	68707.95	41576.85	72021.55	92973.8	NA	NA	NA	88904.66	2064.747	NA	NA
33	P08311	CTSG	NA	NA	46154.89	NA	NA	67879.78	1804.538	NA	53026.19	NA	NA	68927.99	NA	NA	NA	NA	218057.1	78414.15	NA	NA	46895.88	NA	NA	56514.53	66379.24	NA	NA
34	Q9UKU7	ACAD8	46053.91944	31797.32	50179.16	NA	64601.65	NA	75160.02	49228.15	44010.16	28070.84	41974.24	NA	41840.21	NA	42678.39	NA	24335.52	32270.84	46053.92	NA	49467.07	NA	61900.08	NA	2064.747	46053.92	44605.86
35	Q86X76	NIT1	75613.88611	NA	61068.98	63988.55	80199.58	69590.71	1804.538	55745.15	70389.43	NA	84009.8	75506.47	78547.77	84980.21	76153.19	NA	57523.94	40935.27	70713.02	NA	59540.84	70713.02	78753.85	73278.36	55745.15	58932	52415.71
36	P05162	LGALS2	33491.8	NA	35565.03	NA	52415.71	36825.06	1804.538	23560.07	18592.77	NA	36763.92	72761.18	35479.7	50008.51	24907.94	NA	16653.18	22730.31	34916.06	NA	30730.15	NA	32815.68	71139.86	2064.747	NA	25737.06
37	P23946	CMA1	NA	NA	NA	NA	NA	NA	1804.538	NA	NA	NA	NA	NA	NA	NA	NA	NA	61155.07	14049.16	NA	NA	NA	NA	NA	NA	53240.82	NA	NA
38	P01834	IGKC	462133.8694	885197.1	692332.5	484624	296507.9	462133.9	1219164	319228.4	659554.4	351190.2	312295.6	524995.4	566103.9	692332.5	325019.6	494067.2	286640.3	263299.1	499422.8	1130692	706520.3	469971.2	322906.2	438752.3	913960	310472.5	643593
39	P14868	DARS	12567.36389	110112	54554.37	136875.9	30209.1	121022.2	1804.538	114195.5	43350.86	95493.71	29430.84	182241.6	61667.11	201171.9	81193.99	247871.5	161420	94484.9	76929.26	114678.3	54839	177772	50108.55	141996.6	2064.747	95951.08	53026.19
40	Q9H773	DCTPP1	NA	NA	NA	NA	NA	NA	1804.538	46303.49	NA	11589.48	NA	27509.79	NA	NA	NA	26314.17	87070.11	74656.39	NA	NA	NA	NA	NA	NA	2064.747	22251.11	NA

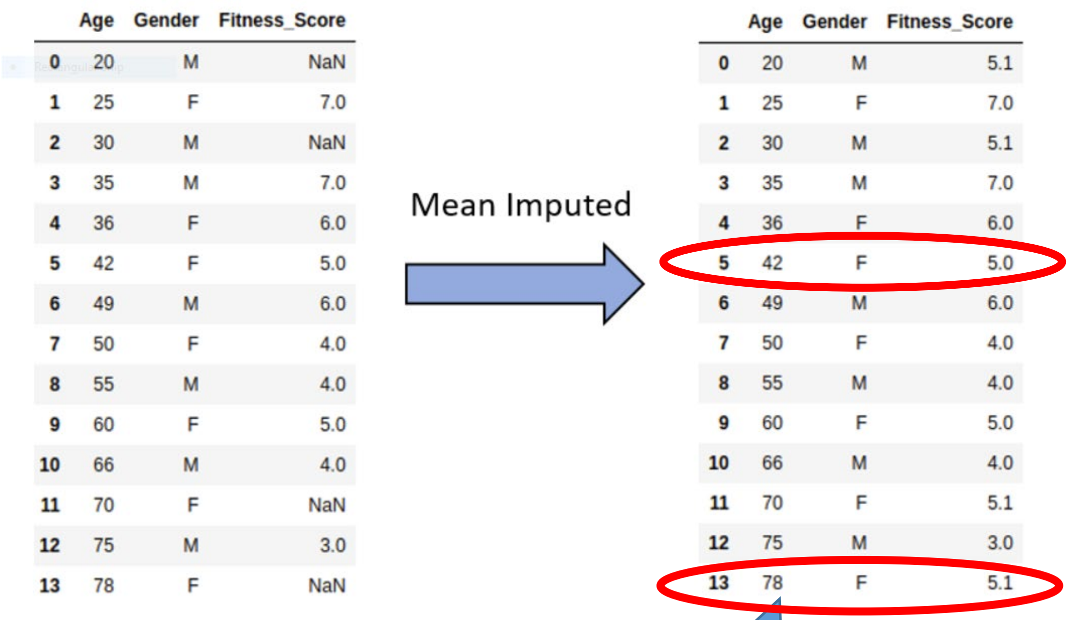
Guo et al. *Nature Medicine*, *21*, 407, 2015

Common missing-value imputation methods

Impute based on the mean value of the corresponding feature

Determine highly correlated variables, impute by regression

Impute based on the mean of k nearest neighbours



	Age	Gender	Fitness_Score
0	20	M	NaN
1	25	F	7.0
2	30	M	NaN
3	35	M	7.0
4	36	F	6.0
5	42	F	5.0
6	49	M	6.0
7	50	F	4.0
8	55	M	4.0
9	60	F	5.0
10	66	M	4.0
11	70	F	NaN
12	75	M	3.0
13	78	F	NaN

Mean Imputed

	Age	Gender	Fitness_Score
0	20	M	5.1
1	25	F	7.0
2	30	M	5.1
3	35	M	7.0
4	36	F	6.0
5	42	F	5.0
6	49	M	6.0
7	50	F	4.0
8	55	M	4.0
9	60	F	5.0
10	66	M	4.0
11	70	F	5.1
12	75	M	3.0
13	78	F	5.1

Mean Imputation of the Fitness_Score

Hmm… this 78-yr-old is as fit as a 42-yr-old?

Image credit: Kacper Kubara

Exercise

You have two batches with lots of missing values

Do you normalize / remove batch effects first, or do you impute missing values first?

Do you combine the two batches and do missing-value imputation on the combined data, or do you do missing-value imputation on the two batches separately?

Why batch-sensitization is impt for missing-value imputation

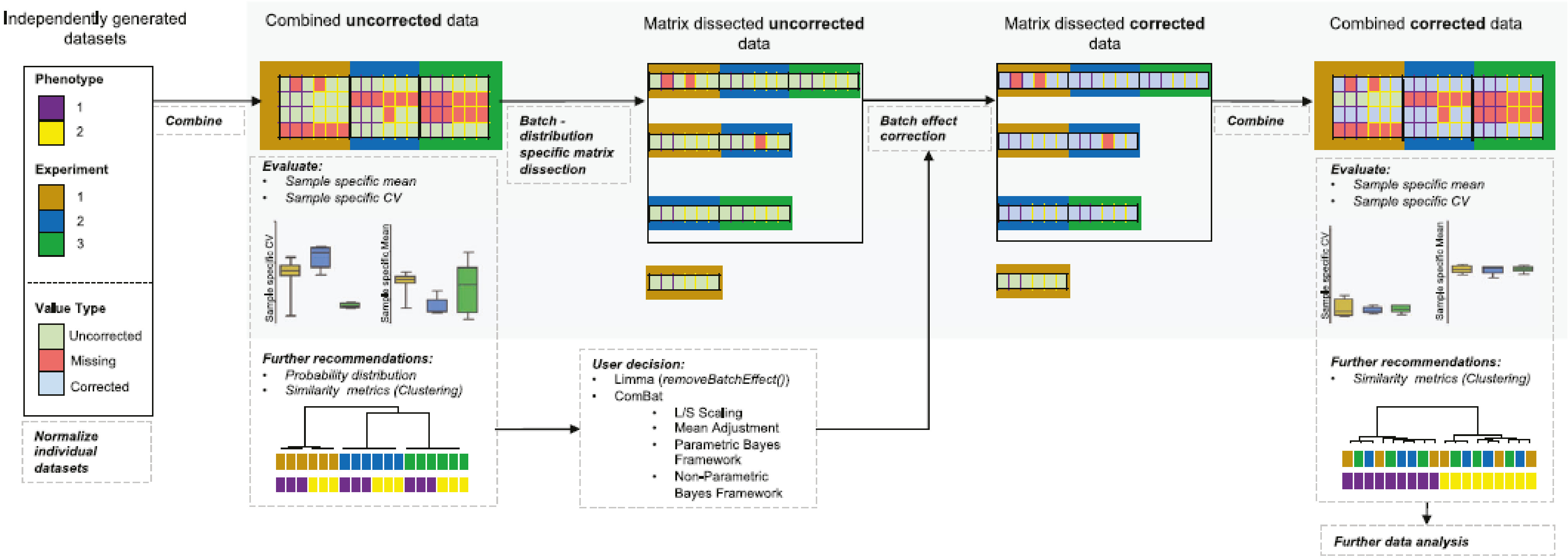
Intentionally left blank

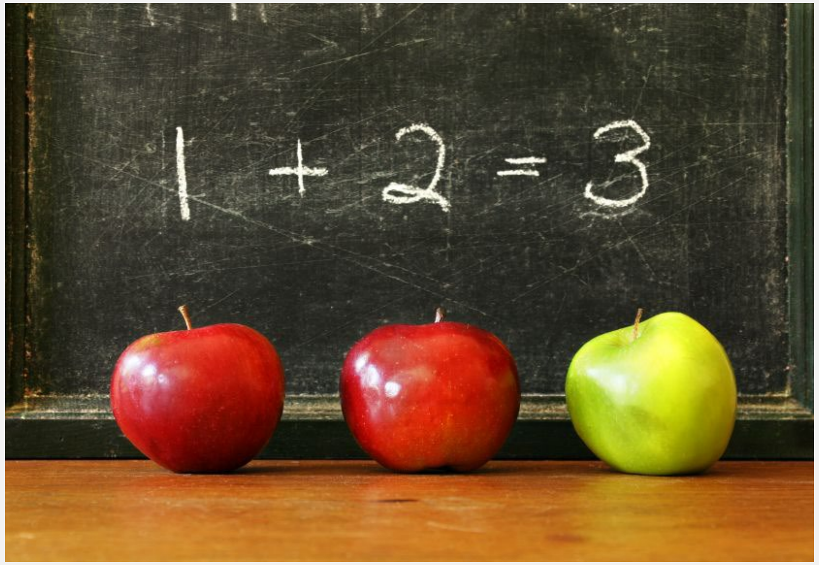
Sun & Goh, https://doi.org/10.21203/rs.3.rs-1328989/v1

HarmonizR

Voss et al., *Nature Communications*, 13:3523, 2022

- Combined data into one matrix
- Extract submatrices w/ few missing values
- Batch correct each submatrix
- Put them back together




1 + 2 = 3

Summary

Batch effects are insidious and unavoidable in omics data

Batch-effect correction can introduce artifacts into data

Missing values are prevalent in some omics data types (e.g., proteomics MS and scRNA-seq)

Missing-value imputation in the presence of batch effects is tricky

Batch-effect correction in the presence of missing values is tricky

References

Leek et al., “Tackling the widespread and critical impact of batch effects in high-throughput data”, *Nat. Rev. Genet.*, 11(10):733-739, 2010

Buttner et al., “A test metric for assessing single-cell RNA-seq batch corrections”, *Nat. Methods*, 16:43-49, 2019 (kBET)

Goh & Wong, “Protein complex-based analysis is resistant to the obfuscating consequences of batch effects – a case study in clinical proteomics”, *BMC Genomics*, 18:142, 2017 (PCA side-by-side boxplot)

Zhao et al., “How to do quantile normalization correctly for gene expression data analysis”, *Scientific Reports*, 10:15534, 2020

Zhou et al., “Examining the practical limits of batch effect-correction algorithms: When should you care about batch effects?”, *J Genet. Genomics*, 46:433-443, 2019