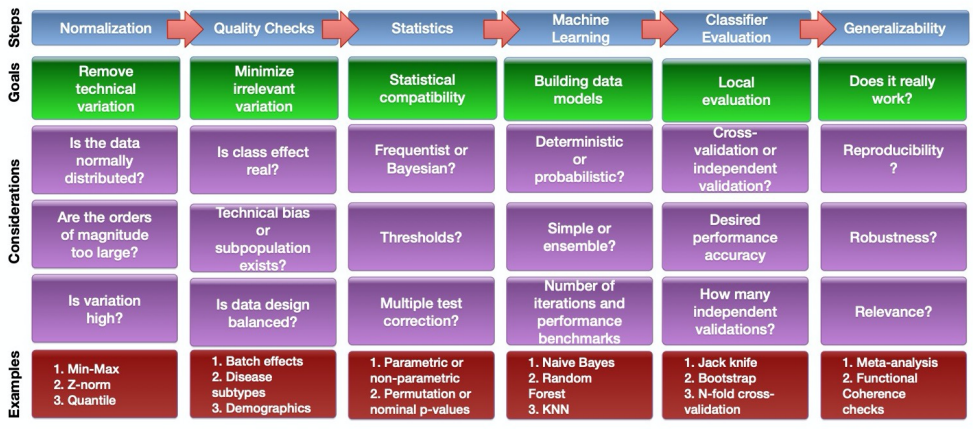
Some (surprisingly) simple hacks to get more out of your data

Wilson Wen Bin GOH (2020)

2

Data analysis is complex

- There are dependencies and assumptions at every step
- Errors generated upstream can be propagated and obfuscated downstream

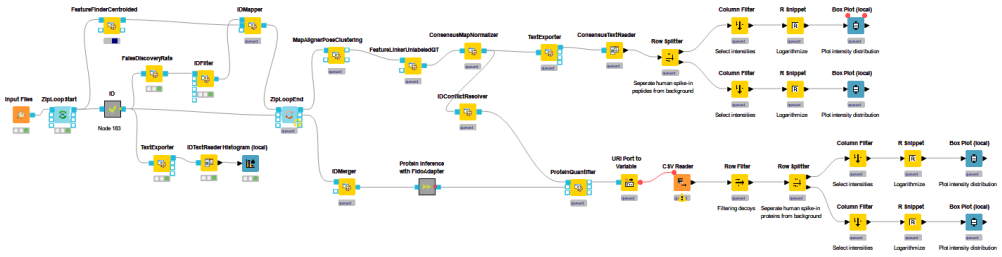


Goh and Wong, *The Birth of Bio-data Science: Trends, Expectations, and Applications. GPB* (2020)

3

Data analysis is complex

- An MS proteomics pipeline built using OpenMS and KNIME by my PhD student Weijia
- Each node is a data processing step, requiring different parameters and settings
- How do you know if your settings are suitable?



4

Talk outline

- **Some surprisingly simple ways to avoid data errors**
  - Missing Value Imputation
  - Normalization
  - Wrong filtering
- **Robustness of network-based analysis**
- **Longitudinal study of mental health risk in Singaporean youth**

5

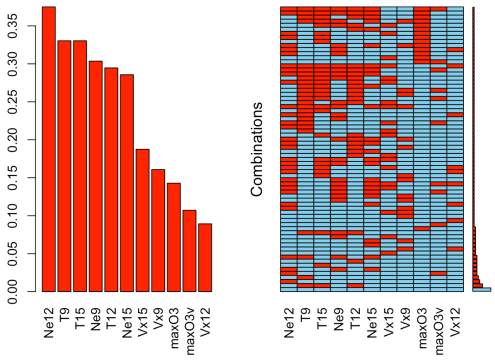
Some surprisingly simple ways to avoid data errors

Missing Value Imputation

Joint work with my CNY-FYP students, Priscila Sun and Wee Yuhui

6

What is Missing Value Imputation (MVI)?

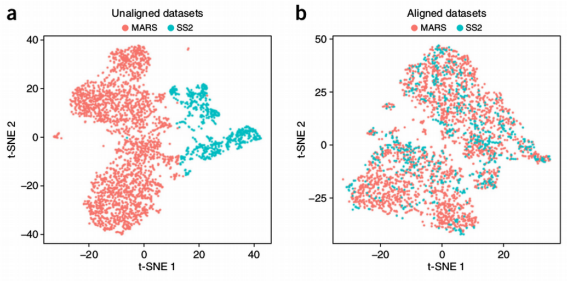


From Jossie J, *Handling missing values with R.* (2018), image source

- MVI is the task of estimating and inserting the value of a missing data value
- Missing values abound in real world data
- Some simple MVI approaches include 0 imputation, or variable-wise global mean imputation
- We will focus on variable-wise global mean imputation

7

What are batch effects?



From Butler et al., *Nature Biotechnology* **36**, p. 411–420 (2018), image source

- Batch effects are technical sources of variation
- Can be due to machine, reagent, experimenter
- Can generate false positives and false negatives
- Exact nature is likely heterogeneous and complex
- Can be estimated and removed via batch effect correction algorithms (BECA)

8

What happens when MVI is done but batch effects exist?

When you have missing values, you tend to impute based on the variable-wise global average, even if a batch factor is known to exist

Or you are "blissfully" unaware of it as you received a fully populated matrix from someone else

Would ignoring the batch factor during Missing value imputation (MVI) confound analysis?

Does it matter for all kind of batch effects?

9

Simulations

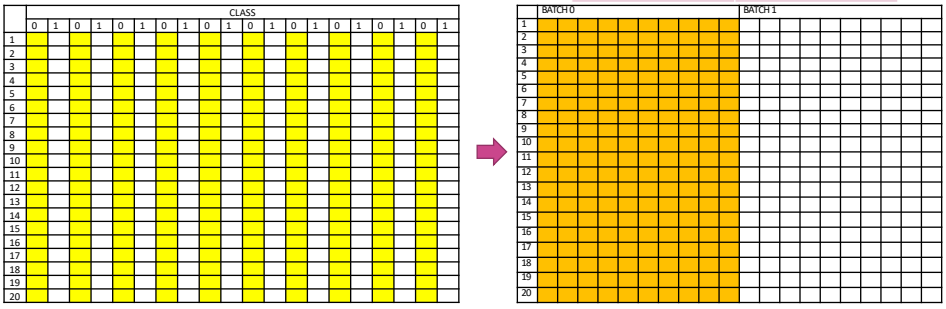
Create class effects

- Class effects loaded on class 0 only
- Multiply class 0 by class factor of 1.5

Create batch effects

- Batch effects loaded on batch 0

Types of batch effects	Formula *where X=data matrix, Z=1.2, Y=sqrt(5)*
Multiplicative	Z(X)
Additive	X+Y
Multiplicative+ Additive	Z(X) + Y

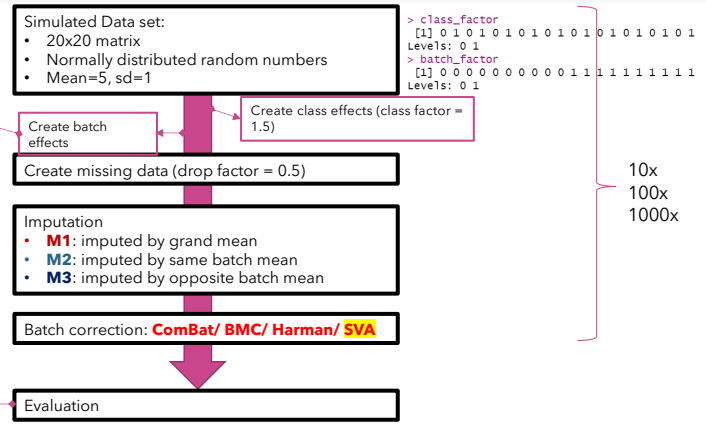


10

Simulation design (at a glance)

Types of batch effects	Formula *where X=data matrix, Z=1.2, Y=sqrt(5)*
Multiplicative	Z(X)
Additive	X+Y
Multiplicative + Additive	Z(X) + Y

Feature Selection	Score
Batch Correction	• gPCA delta
Imputation Accuracy	• RMSE • Statistical testing



11

What the labels mean

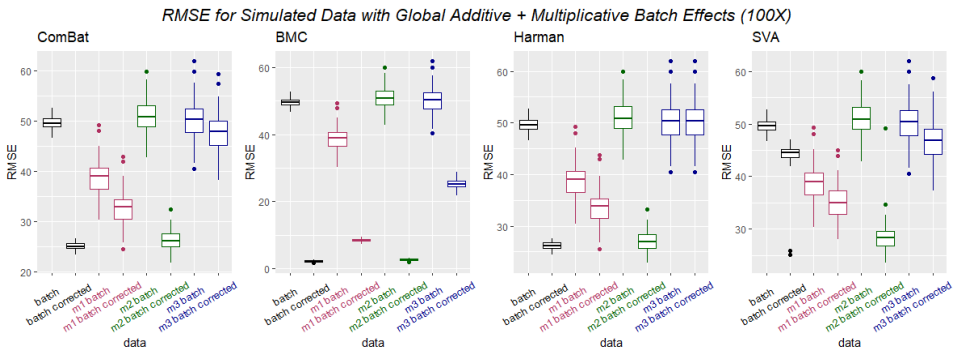
Legend	gPCA delta Scores of …
True null	data + class effects
batch	data + class effects + batch effects
batch corrected	data + class effects + batch effects + batch effect correction
m1 batch	data + class effects + batch effects + 50% missing data + m1 (grand mean) imputation
m1 batch corrected	data + class effects + batch effects + 50% missing data + m1 (grand mean) imputation + batch effect correction
m2 batch	data + class effects + batch effects + 50% missing data + m2 (same batch mean) imputation
m2 batch corrected	data + class effects + batch effects + 50% missing data + m2 (same batch mean) imputation + batch effect correction
m3 batch	data + class effects + batch effects + 50% missing data + m3 (opposite batch mean) imputation
m3 batch corrected	data + class effects + batch effects + 50% missing data + m3 (opposite batch mean) imputation + batch effect correction

Legend	RMSE Scores
batch	(data + class effects + batch effects) - **(data + class effects)**
batch corrected	(data + class effects + batch effects + batch effect correction) - **(data + class effects + batch effect correction)**
m1 batch	(data + class effects + batch effects + 50% missing data + m1 (grand mean) imputation) - **(data + class effects)**
m1 batch corrected	(data + class effects + batch effects + 50% missing data + m1 (grand mean) imputation + batch effect correction) - **(data + class effects + batch effect correction)**
m2 batch	(data + class effects + batch effects + 50% missing data + m2 (same batch mean) imputation) - **(data + class effects)**
m2 batch corrected	(data + class effects + batch effects + 50% missing data + m2 (same batch mean) imputation + batch effect correction) - **(data + class effects + batch effect correction)**
m3 batch	(data + class effects + batch effects + 50% missing data + m3 (opposite batch mean) imputation) - **(data + class effects)**
m3 batch corrected	(data + class effects + batch effects + 50% missing data + m3 (opposite batch mean) imputation + batch effect correction) - **(data + class effects + batch effect correction)**

12

Imputing based on M2 has the lowest RMSE across any BECA

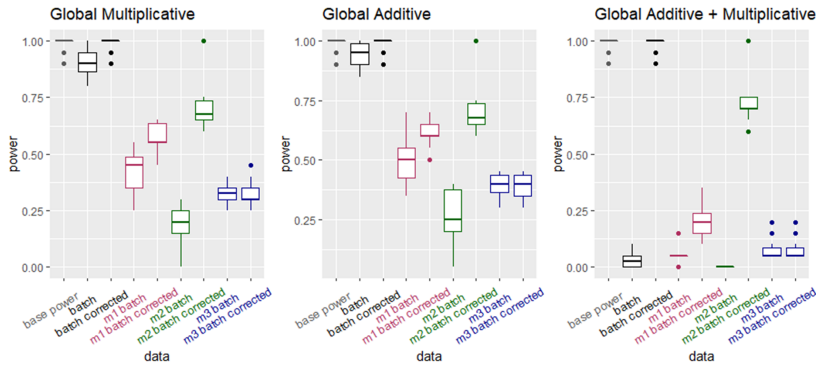
- The lower the RMSE the better
- It does not matter which BECA you use
- You should impute based on M2



13

Low RMSE also translates to good power

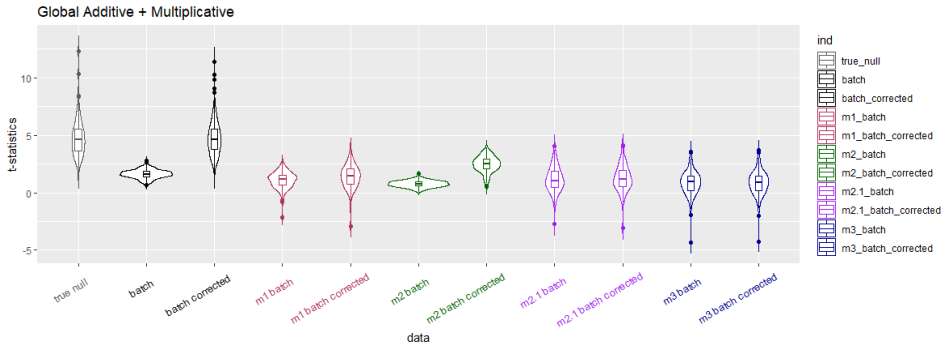
- Power is the proportion of correct genes returned
- M2 gives the highest power
- However, it never performs as well as batch corrected (without missing values)



14

But beware the drop in effect size estimations

- The t-statistic is an estimate of effect size
- Note that post MVI distributions are much flatter than the original
- M2.1 is a special scenario considering both batch and effect factors (not discussed)



15

Real data

Types of batch effects	Formula *where X=data matrix, Z=1.2, Y=sqrt(5)*
Multiplicative	Z(X)
Additive	X+Y
Multiplicative+ Additive	Z(X) + Y

Feature Selection	Score
Batch Correction	• gPCA delta • PCA scatterplots • Compare change in medians of sample boxplots
Imputation Accuracy	• RMSE

Genomic Dataset GDS4056, GDS4057:
- Take ER+ samples from GDS4056
  - Dimension = 22,283 features x 32 samples
- Take ER+ samples from GDS4057
  - Dimension = 22,283 features x 42 samples
  - Drop last 10 samples to make batch size even
- 32 samples in batch 4056, 32 samples in batch 4057
- Only one true sample class: ER+

Data pre-processing:
- Z-transform: mean=0, sd=1
- Add 5 to all the data
- Amplify the **batch effect** of the first 32 columns

Create missing data (drop factor = 0.5)

Imputation
- **M1**: imputed by grand mean
- **M2**: imputed by same batch mean
- **M3**: imputed by opposite batch mean

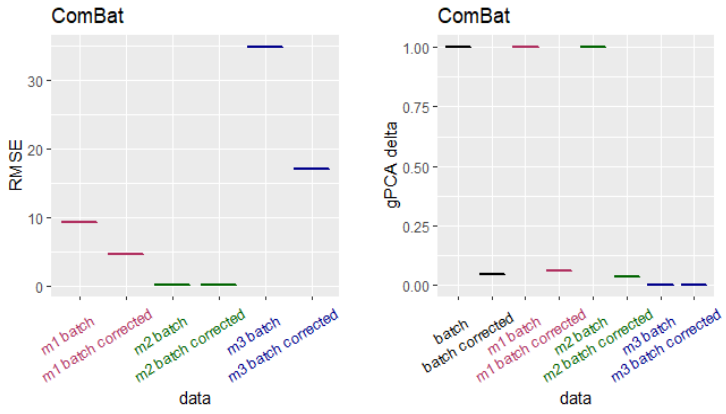
Batch correction: **ComBat**

Evaluation

16

M2 works (beautifully) on genomic data

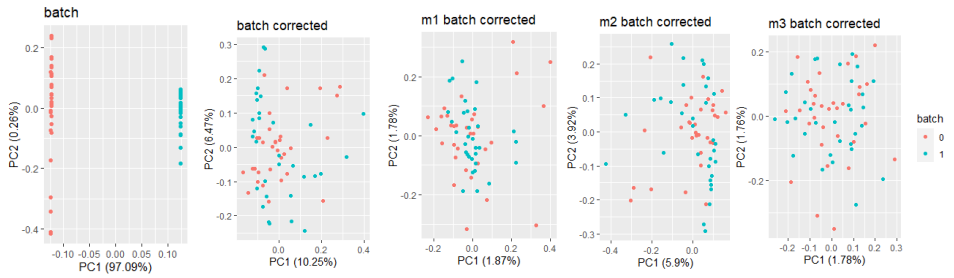
- M2 has the lowest RMSE
- M2 also has a pretty good gPCA delta (a relative estimation of batch proportion in data)



17

M2 works (beautifully) on genomic data

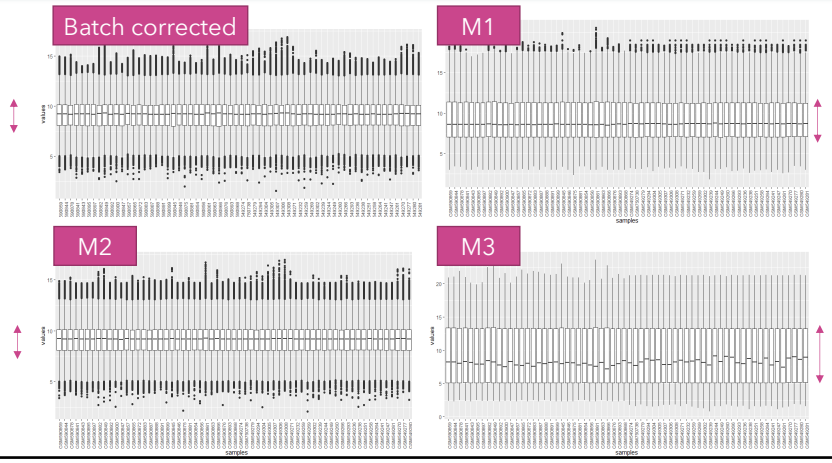
- M2 has comparable gPCA to M1
- In PCA, it also does not appear impressive…but…



18

M2 works (beautifully) on genomic data

- …M1 and M3 do result in increased noise in the data even if the batch effects appear to be "mitigated" (y-axis: sample value distribution)
- Similar findings also for proteomics data (not shown)



19

Where are we on this now?

- Results based on Harman, SVA and BMC are comparable. ComBat is overall best (so we focus on this)
- This is an early result --- we are going to rework this simulations to also include false positives
- Simulate not just MCAR, but also MNAR MVs
- Preliminary data suggests the reduction in power is due to inflated variances -> but have to work out the mathematical proofs as to why this is the case

20

Key Takeaway

If you know a batch factor exists in your data and you want to impute missing values…

…Do not impute based on global mean. Make sure you impute based on same batch samples only...

21

(Some) relevant lab publications

- *Zhou LJ, Sue ACH, **Goh WWB**. Examining the practical limits of batch effect correction algorithms: When should you care about batch effects? Journal of Genetics and Genomics, 46(9):433-443, Sep 2019
- **Goh WWB**, Wong LS. Dealing with confounders in -omics analysis. Trends in Biotechnology, 36(5):488-498, May 2018 ##
- **Goh WWB**, Wang W, Wong LS. Why batch effects matter in omics data, and how to avoid them. Trends in Biotechnology, S0167-7799(17)30036-7, Mar 2017 ##
- **Goh WWB**, Wong LS. Protein complex-based analysis is resistant to the obfuscating consequences of batch effects --- A case study in clinical proteomics. BMC Genomics, 18(Suppl 2):142, Mar 2017 #

22

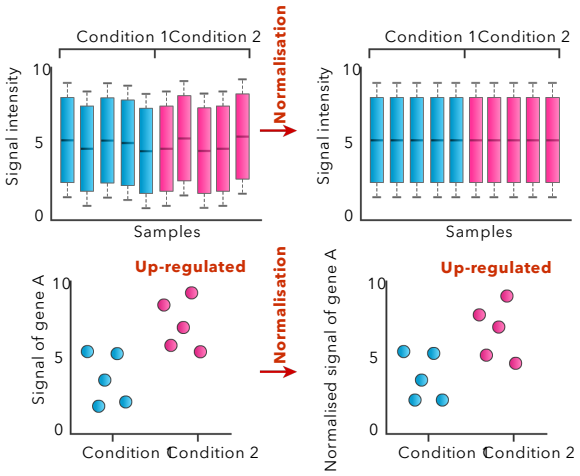
Some surprisingly simple ways to avoid data errors

Normalization

Joint work with my MSc student, Yaxing Zhao and my CNY-FYP student, Joan Jong

23

What is normalization?



Source: Wu et al. 2014

- Normalization is the statistical practice of aligning your samples so that they are cross-comparable
- Normalization works well if two sets of distributions are not too different from each other.
- The common assumptions for normalization are reasonable if similar global signal distributions are seen in the different conditions. In such cases, normalization has little influence on the interpretation of expression data.

24

Quantile normalization (QN) is a commonly used technique

Quantile normalisation is a technique for making two distributions identical in statistical properties.

Raw data			
2	4	4	5
5	14	4	7
4	8	6	9
3	8	5	8
3	9	3	5

Order values within each sample (or column)			
2	4	3	5
3	8	4	5
3	8	4	7
4	9	5	8
5	14	6	9

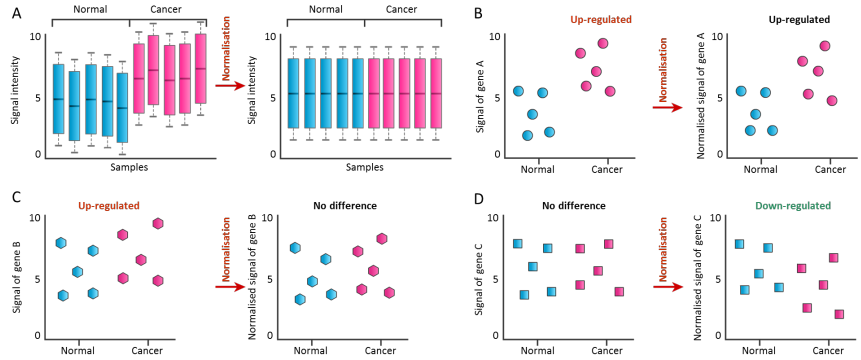
Average across rows and substitute value with average			
3.5	3.5	3.5	3.5
5.0	5.0	5.0	5.0
5.5	5.5	5.5	5.5
6.5	6.5	6.5	6.5
8.5	8.5	8.5	8.5

Re-order averaged values in original order			
3.5	3.5	5.0	5.0
8.5	8.5	5.5	5.5
6.5	5.0	8.5	8.5
5.0	5.5	6.5	6.5
5.5	6.5	3.5	3.5

25

But QN does not work well when sample distributions are strongly different

- A: Strong sample distribution difference
- B: Effect reduction
- C: Effect neutralization
- D: False effects

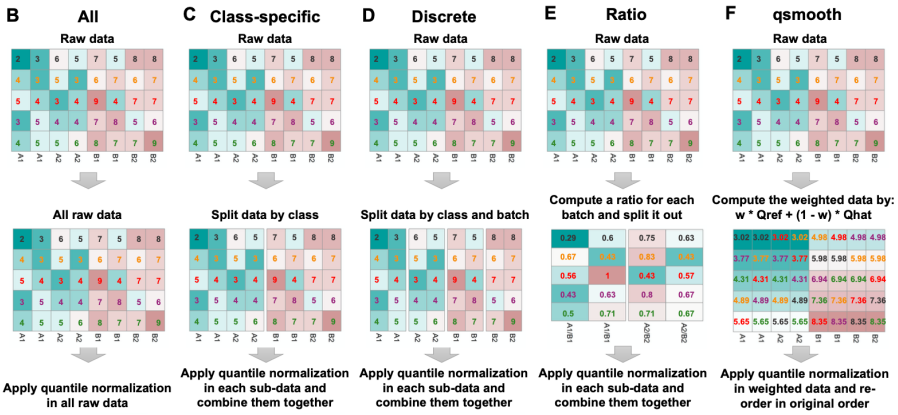


Source: Wu et al. 2014

26

So why not QN each class separately?

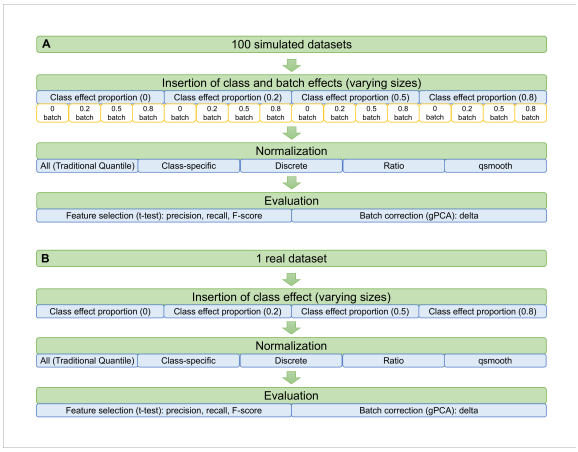
- You can still do quantile normalization
- But you account for other factors in your data
- My proposed winning idea is to simply split data by class labels and QN the splits independently (**Class-specific**)



Yaxing Zhao et al. How to do quantile normalization correctly for gene expression data analysis. *Scientific Reports*, 2020

27

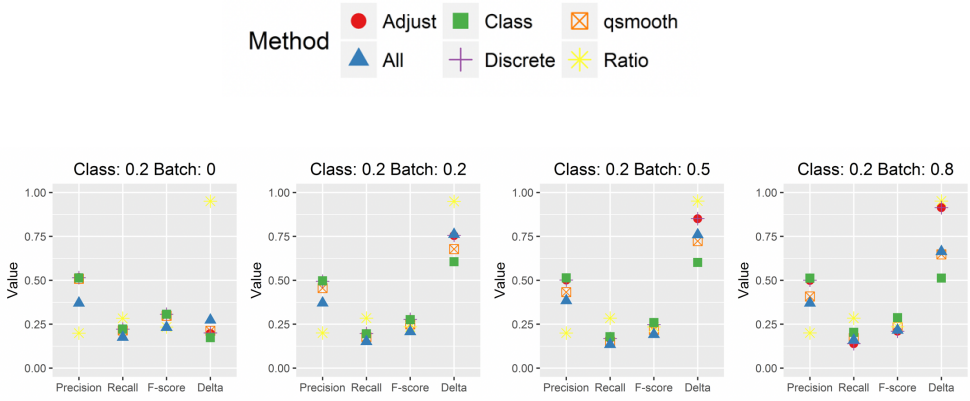
Simulation design (at a glance)



- A: Purely simulated data with differing class and batch effects
- B: Real data with natural batch effects

28

CS-Quantile performs well on simulated data



29

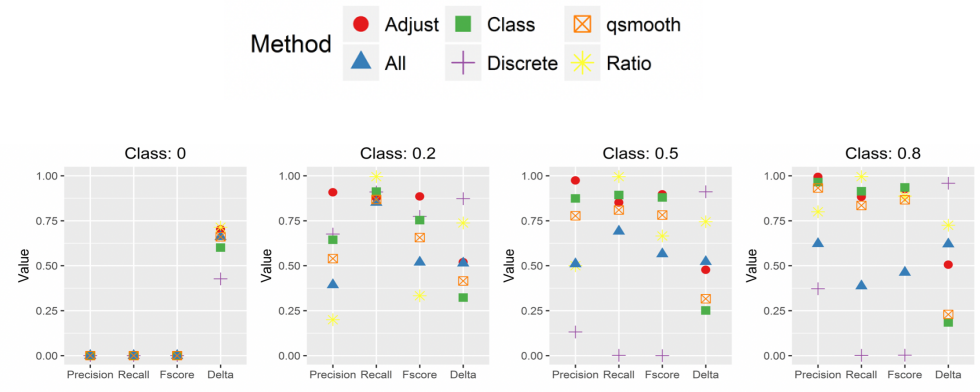
CS-Quantile performs well on simulated data

		F-score				Delta			
Batch		0	0.2	0.5	0.8	0	0.2	0.5	0.8
CEP: 0.2	Adjust	2.5	2.5	2	4	2.5	4	4.5	4.5
	Class-specific	2.5	2.5	2	1	1	1	1	1
	qsmooth	2.5	2.5	5	4	4	2	2	2
	All	5.5	6	5	4	5	4	3	3
	Discrete	2.5	2.5	2	4	2.5	4	4.5	4.5
	Ratio	5.5	5	5	4	6	6	6	6

Shown are ranks where 1 is the best

30

CS-Quantile performs also well on real data



Cannot vary batch effects but can introduce class effects of various sizes

31

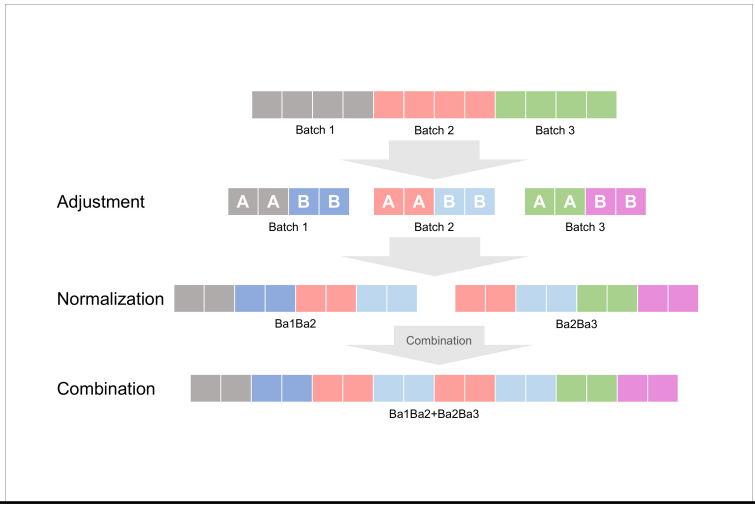
CS-Quantile performs also well on real data

	F-score				Delta			
CEP	0	0.2	0.5	0.8	0	0.2	0.5	0.8
Adjust	0	1	1	2	5	3.5	3	3
Class-specific	0	3	2	1	2	1	1	1
qsmooth	0	4	3	3.5	3	2	2	2
All	0	5	5	5	4	3.5	4	4
Discrete	0	2	6	6	1	6	6	6
Ratio	0	6	4	3.5	6	5	5	5

Shown are ranks where 1 is the best

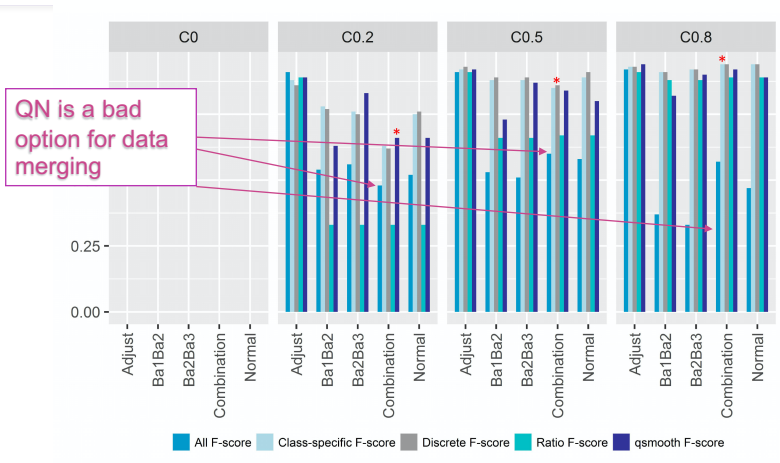
32

Simulating data merging



33

You can't combine data well using the usual QN approach ("All")



34

CS-Quantile gives very good feature selection while also reducing batch effects

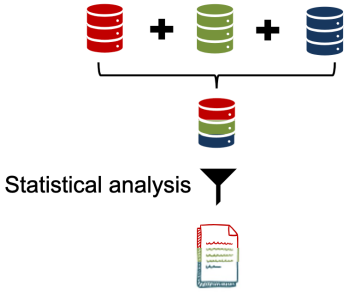
	F-Score			Delta		
CEP	0.2	0.5	0.8	0.2	0.5	0.8
All	4	5	5	4	4	4
Class-specific	2	2	1	2	1	1
Discrete	3	1	2	1	2	3
Ratio	5	4	4	5	5	5
qsmooth	1	3	3	3	3	2

Ranks are shown in table --- where 1 is the best and 5 is the worst

35

CS-Quantile is suitable for mega-analysis

- We combine 4 datasets together for serial analysis (aka mega analysis)
- Batch and class effects exist
- Amongst methods for normalization and batch effect removal, which ones do the best?



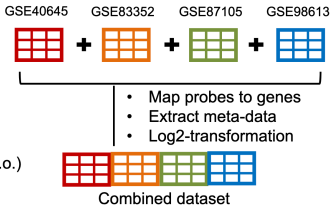
Mega-analysis

Data mining, cleaning and mapping

4 batches (GEO datasets) and 2 classes (old & young)

Inclusion Criteria

- Gene expression array
- Vastus lateralis tissue
- Illumina HumanHT-12 platform
- Has young and old samples (<= and > 50 y.o.)



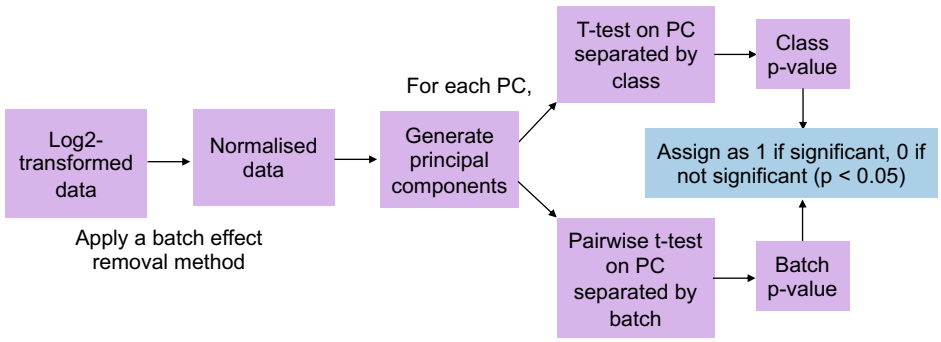
36

Combinatorial methods examined

Method	Normalisation	BECA	
1	-	-	No further normalisation
2	Quantile Normalisation (QN)	-	Normalisation or BECA
3	Z-Normalisation	-	
4	-	ComBat	
5	Class-specific QN	-	
6	Quantile Normalisation (QN)	ComBat	Combinations of normalisation and BECA
7	Z-Normalisation	ComBat	
8	Class-specific QN	ComBat	

37

Determination of efficacy via PCA



38

Determination of efficacy via PCA

PC	Class	Batch
1	0	1
2	0	1
3	0	1
4	0	0
…	…	…
59	1	0

All top 3 PCs correlated with batch – batch effect is dominant

PCs correlated with class ranked low – biological variation is very subtle

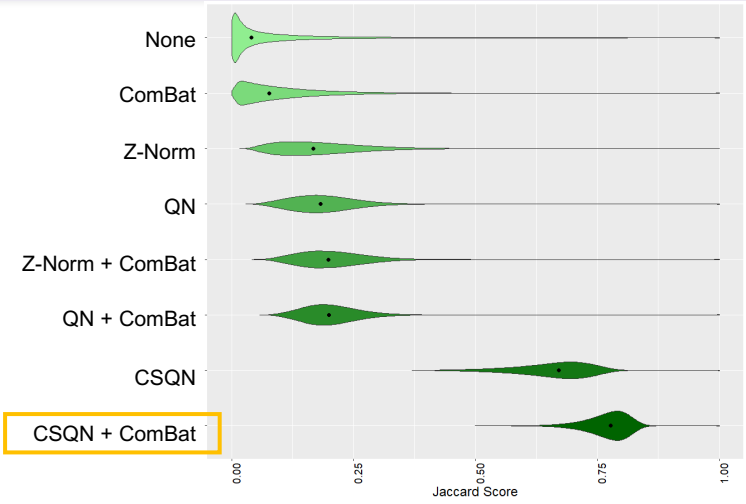
39

CS-Quantile + ComBat is a powerful combination…improving signal-to-noise

Index	Normalisation	BECA	Top 3 PCs Correlated with Class?	Top 3 PCs Correlated with Batch?
1	-	-	NO	YES
2	Quantile Normalisation (QN)	-	NO	YES
3	Z-Normalisation	-	NO	YES
4	-	ComBat	NO	NO
5	Class-specific QN	-	YES	YES
6	Quantile Normalisation (QN)	ComBat	NO	NO
7	Z-Normalisation	ComBat	NO	NO
8	Class-specific QN	ComBat	YES	NO

40

CS-Quantile + ComBat is a powerful combination…and high reproducibility



$$Jaccard\ Score = \frac{|A \cap B|}{|A \cup B|}$$

The Jaccard Score is a measure of pairwise similarity. The higher it is, the more similar the independent outcomes are, signifying reproducibility

41

Where are we on this now?

CS-Quantile is recently published in Scientific Reports!

We are further investigating its usage in mega-analysis, class-batch imbalances and other practical deployments

42

Key Takeaway

IF YOU ABSOLUTELY MUST USE QN, CONSIDER USING CS-QUANTILE INSTEAD…

AND IF YOU WANT TO JOIN DATA IN SERIES, CONSIDER CS-QUANTILE WITH COMBAT.

43

(Some) relevant lab publications

- **Zhao Y, Wong LS, **Goh WWB**. How to do quantile normalization correctly for gene expression data analyses. Scientific Reports, Accepted
- **Goh WWB**, Wong LS. Turning straw into gold: how to build robustness into gene signature inference. Drug Discovery Today, 24(1):31-36, Jan 2019 ##
- **Goh WWB**, Wong LS. Breast Cancer Signatures Are No Better Than Random Signatures Explained. Drug Discovery Today, 23(11):1818-1823, Nov 2018 ##
- **Wang W, Sue ACH, **Goh WWB**. Null-hypothesis statistical testing in clinical proteomics: With great power comes great reproducibility. Drug Discovery Today, 22(6):912-918, June 2017 ##
- **Wang W, **Goh WWB**. Sample-to-sample p-value variability and its implications in multivariate analysis. International Journal of Bioinformatics Research and Applications, 14(3):235-254, Jan 2018

44

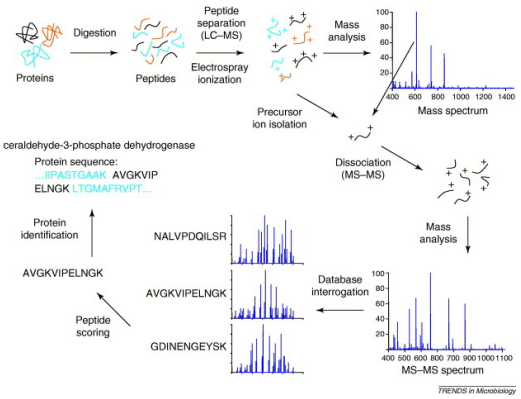
Some surprisingly simple ways to avoid data errors

Wrong filtering

Joint work with my URECA student, Bertrand Wong and PhD students, Kong Weijia and Longjian Zhou

45

What is the two-peptide rule?

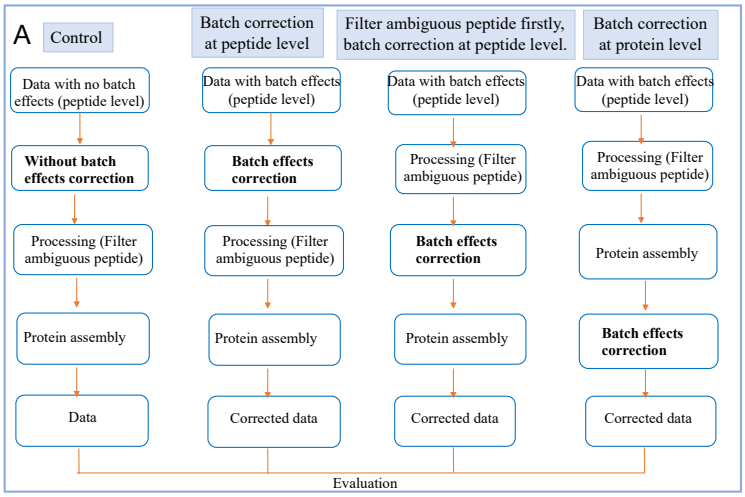


- In proteomics, peptides are inferred from mass spectra
- Peptides are short amino acid segments
- A peptide can be mapped uniquely to a protein
- A peptide that maps to more than 1 protein is known as ambiguous
- Positive identification of a protein requires at least 2 uniquely mapped peptides
- All other peptides are discarded
- **Why is this problematic?**

46

There is useful signal locked in ambiguous peptides

Batch inference

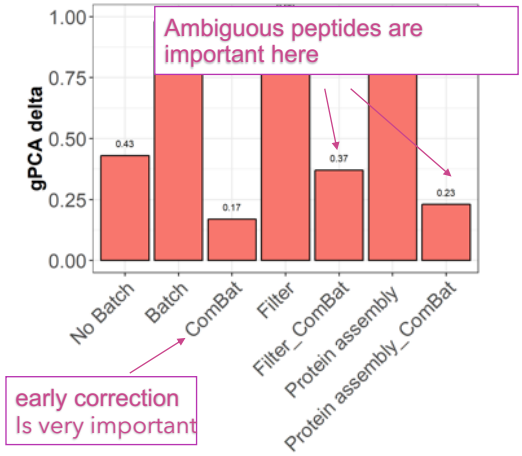


47

There is useful signal locked in ambiguous peptides

Using a kidney tissue dataset from the Aebersold group

- ***"No Batch"*** *--- no batch effects present*
- ***"Batch"*** *--- batch effects present in the peptide data matrix*
- ***"Combat"*** *--- This is the result of batch correction at the peptide level*
- ***"Filter"*** *--- Estimated batch effects after removal of ambiguous peptides*
- ***"Filter_ComBat"*** *--- Remaining batch effect post "Filter"*
- ***"Protein assembly"*** *--- Batch effects present in the assembled protein data matrix.*
- **"Protein assembly_ComBat"** --- Remaining batch effects post "protein assembly".



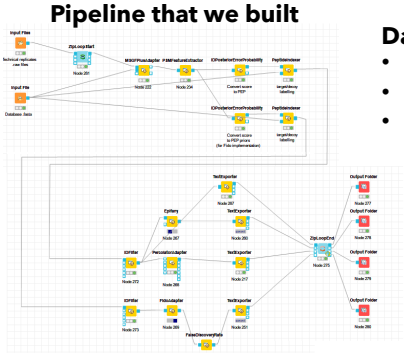
gPCA is an estimator of batch effects

48

There is useful signal locked in ambiguous peptides

Proteome coverage

Pipeline that we built



Data: IPRG2016

- Inferring Proteoforms from Bottom-up Proteomics Data
- Especially useful for identifying ambiguous peptides
- 3 data sets (B1, B2 and B3) --- assay reproducibility

	False Discovery Rate (Verified/All reported)		
	B1	B2	B3
Without inference	372/762(48.8 %)	370/729(50. 8%)	369/745(49.5 %)
Epifany	359/509(71.4 %)	358/462(77. 5%)	357/472(75.6 %)
Percolator	372/762(48.8 %)	370/729(50. 8%)	369/745(49.5 %)
Fido	352/363(97.0 %)	353/366(96. 5%)	353/363(97.2 %)

Percolator "ignores" ambiguous peptide information

49

We propose replacing the two-peptide rule by the following three requirements for reporting a protein P

The first requirement is same as HPP's; viz. the set of reported PSMs (i.e. matched peptides) achieves a global FDR ≤ 1%, and each of the matched peptide achieves a local FDR ≤ 10%.

The second requirement is that there is a set Ω of non-nested matched peptides that supports reporting P with Prob(P is a false report | Ω) ≤ 1%.

The third and last requirement is that P is the only protein (in the reference database) that contains all the peptides in Ω.

Note that there is no requirement for any peptide in Ω to be unique to P.

50

Where are we on this now?

We are developing a new protein assembly algorithm using both ambiguous information and attributable quantitation information

We are looking at how early batch effect removal, at spectra level, can lead to greatly improved proteome coverage and quantitation

51

Key Takeaway

THERE IS VALUE IN AMBIGUOUS PEPTIDES. DO NOT SIMPLY DISCARD

THE TWO-PEPTIDE RULE IS VERY DATED AND VERY CONSTRAINING

ADHERE TO IT AND YOU WILL NOT ONLY LOSE AMBIGUOUS PROTEINS, YOU WILL ALSO REDUCE YOUR DIRECT PROTEOME COVERAGE

Beyond proteomics, check whether those filtering rules you are using now make sense. Do not just blindly follow. Or…check with a bioinformatician/statistician/data scientist.

52

(Some) relevant lab publications

- **Goh WWB**, Wong LS. Spectra-first feature analysis in clinical proteomics --- A case study in renal cancer. Journal of Bioinformatics and Computational Biology, 14(5):1644004, Oct 2016
- **Goh WWB**, Wong LS. Advanced bioinformatics methods for practical applications of proteomics. Briefings in Bioinformatics, 20(1):347-355, Jan 2019 ##
- **Goh WWB**, Wong LS. Computational proteomics: Developing a robust analytical strategy. Drug Discovery Today, 19(3):266-274, Mar 2014 ##

53

Robustness of network-based analysis

Joint work with my intern Damien Chua and MSc Student, Yaxing Zhao

54

What is network-based analysis?

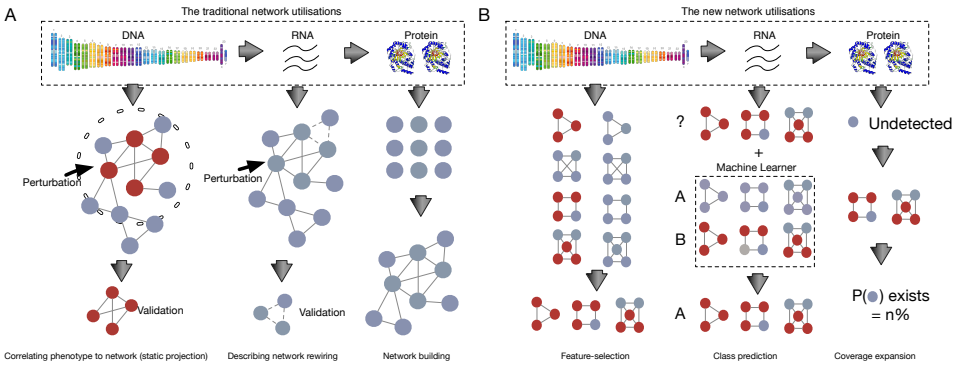
Network-based analysis is an umbrella term describing the integration of molecular information against a systems-based context

Many of these techniques are descriptive and non-statistical

55

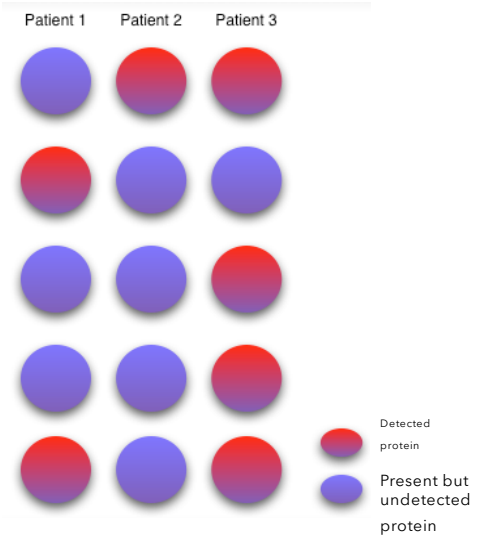
What network-based analysis should be

- Instead of merely describing how some genes fall onto some network
- You can devise scoring functions on networks
- For performing statistical feature selection, machine learning prediction and missing protein prediction



Goh & Wong, Integrating Networks and Proteomics: Moving Forward. Trends in Biotechnology, 2016.

56



Using protein complexes to enhance proteomics --- Basic ideas

- Suppose the failure to form a protein complex causes a disease
- *If any component protein is missing, the complex can't form*
- Diff patients suffering from the disease can have a diff protein component missing
- *Construct a profile based on complexes?*

57

Postulate: Chance of a protein complex being present approx fraction of its member proteins being reported in the screen

Say a proteomics screen has 75% reliability; {A, B, C, D, E} is a complex; and screen reports A, B, C, D only

The complex has 60% (= 0.75 * 4 / 5) chance to be present

E has >60% chance to be present, as presence of complex implies presence of its constituents ... improving coverage

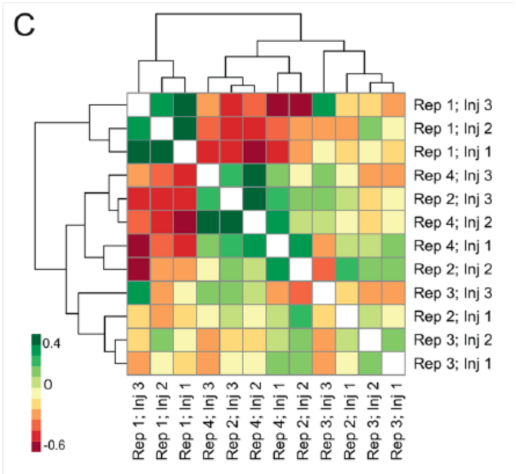
A, B, C, and D each has 90% (= 100% * 0.6 + 75% * 0.4) chance of being present, i.e. >75% ... removing noise

Using protein complexes to enhance proteomics --- Basic ideas

- And extend this concept into some mathematical conceptualization
- We focus on human complexes (of size at least 5) from CORUM are used as reference complexes
- Other networks such as pathways and clusters from protein interaction networks can also be used (but present other problems)

58

Improved consistency in proteomic profile analysis



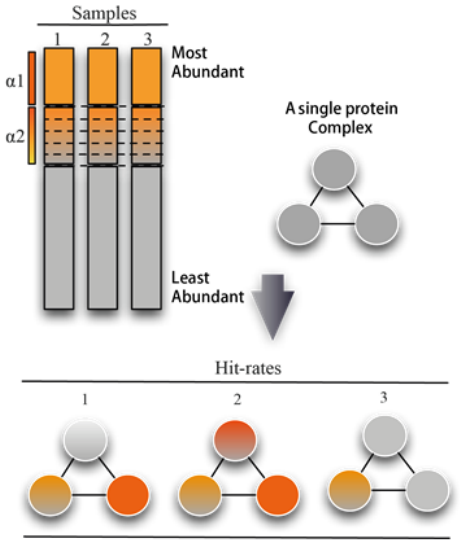
Proteomic profiles sometimes not sufficiently consistent

- A human kidney tissue, digested in quadruplicates, analyzed in triplicates
- Guo et al. *Nature Medicine*, 21(4):407-413, 2015
- Correlation betw replicates is not good (~0.4)
- Technical replicates of the same biological replicate are not tightly clustered

59

qPSP: Constructing a signature profile based on protein complexes

1) In a sample, assign fuzzy score to proteins
2) Hit rate of a complex C wrt a sample S is sum of the wt of proteins in C in S
   score(C,Si) = $\Sigma p \in C$ fs(p,Si) / |C|
3) Complex C is significant if {score(C,S_i) | $S_i \in A$} is very different by t-test from {score(C,S_i) | $S_i \in B$}

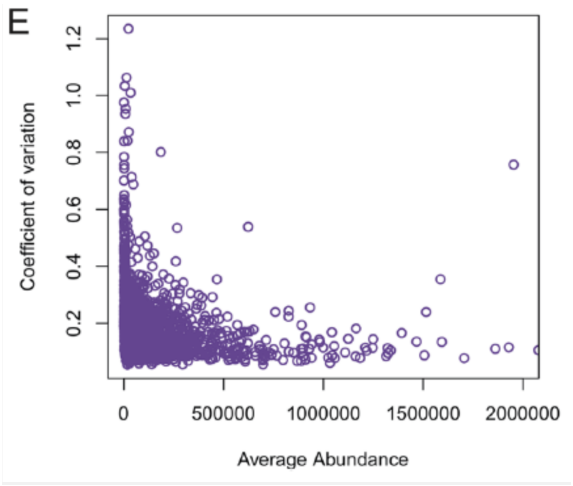


Goh et al. Quantitative proteomics signature profiling based on network contextualization. *Biology Direct*, 2015

60

Why use fuzzy scores?

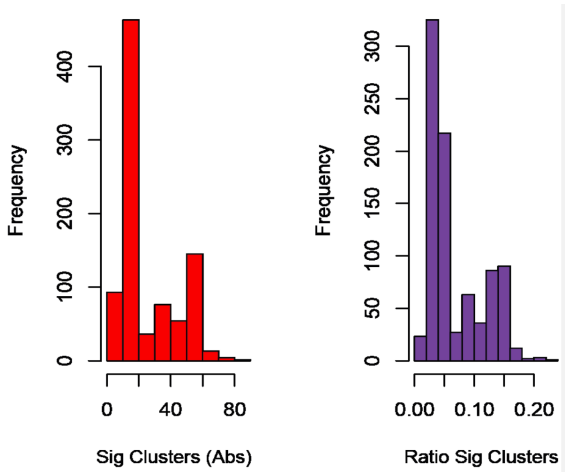
- Low-abundance proteins have very high coefficient of variation
- They are very noisy
- Fuzzy scoring mitigates this
- Fuzzy scoring in itself is a powerful normalization method known as Gene Fuzzy Scoring (GFS)



61

False-positive rate analysis

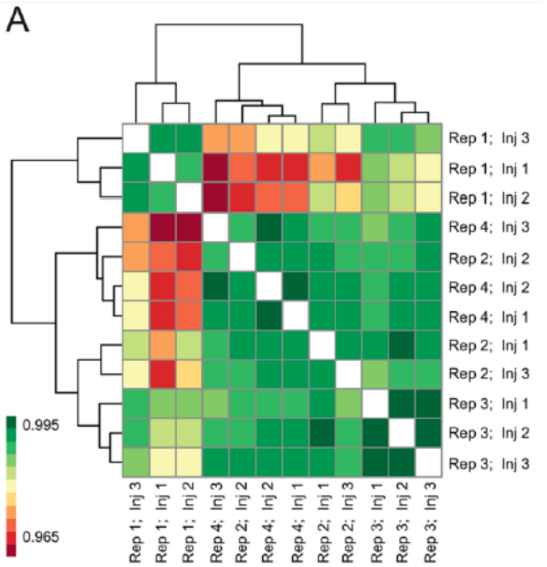
- 12 kidney controls randomly assigned into two groups of equal size, and qPSP performed many rounds
- False-positive rate well within the expectation levels
- *Sig Clusters (Abs)*
Exp: 19, Observed: 16
- *Sig Clusters (Ratio)*
Exp: 5%, Observed: 4%



62

Consistency of qPSP

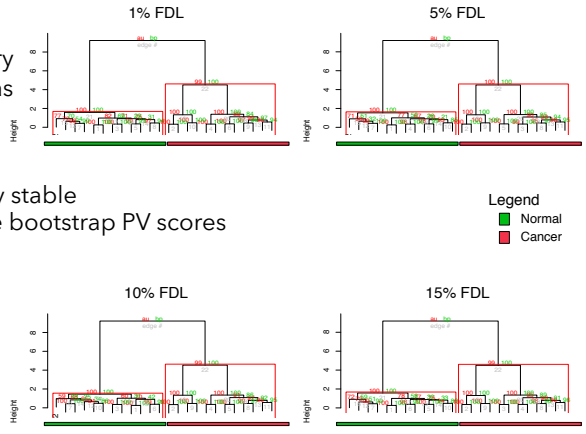
- Clustering of benchmarking control data based on protein complexes (i.e. qPSP)
- Correlation betw replicates is >0.95
- *Cf. 0.4 based on proteins*
- Also, technical replicates are better clustered



63

High noise resistance in real samples

FDL:
False Discovery Levels (same as FDR)



Clusters are very stable
As shown by the bootstrap PV scores

Goh et al. Quantitative proteomics signature profiling based on network contextualization. *Biology Direct*, 2015

64

Application to renal & colorectal cancers

- A: Clustering for renal cancer
- B: Clustering for colorectal cancer
- Dendrograms show discrimination between sample classes is very strong and stable

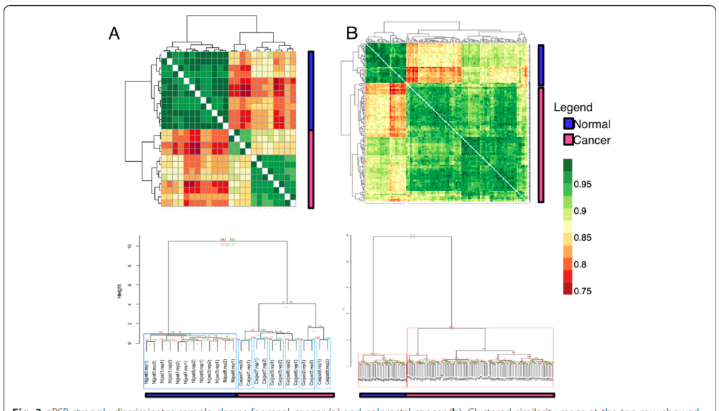
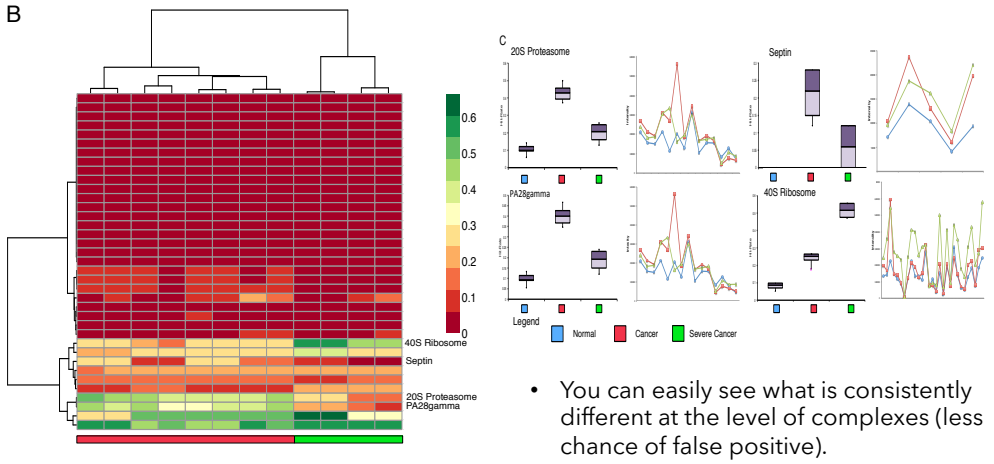


Fig. 3 qPSP strongly discriminates sample classes for renal cancer (**a**) and colorectal cancer (**b**). Clustered similarity maps at the top row showed specific and consistent segregation of non-cancer and cancer samples. The trees below the heatmaps are from bootstrap analysis (PVCLUST), which demonstrates that the discrimination between sample classes based on qPSP hit-rates is highly stable

65

Less noise…easier to analyze



- You can easily see what is consistently different at the level of complexes (less chance of false positive).
- The hit-rates are a useful data summary

Goh et al. Quantitative proteomics signature profiling based on network contextualization. *Biology Direct*, 2015

66

Further improving consistency in proteomic profile analysis via Paired Fuzzy Subnets (PFSNET)

1.
2. **Fuzzy scoring** --- eliminate noise and erroneous measurements
3. **Class weighting** --- Signal boosting
4. **Network scoring** --- Signal boosting

67

Further improving consistency in proteomic profile analysis via Paired Fuzzy Subnets (PFSNET)

- Improves signal detection by weighting on class-proportion support

Statistical testing

- **Re-designed** (you have more information due to the networks although the degrees-of-freedom remain the same
- **Robustness** --- the p-value (empirical) is calculated using Fisher's permutation test instead

Goh & Wong. **Advancing clinical proteomics via analysis based on biological complexes: A tale of five paradigms**. *Journal of Proteome Research*, 15(9):3167--3179, July 2016.

68

Further improving consistency in proteomic profile analysis via Extremely Small SubNets (ESSNET)

- Null hypothesis is "Complex C is irrelevant to the diff betw patients and normals, and the proteins in C behave similarly in patients and normals"
- No restriction to most abundant proteins
- Potential to reliably detect low-abundance but differential proteins

Let g_i be a protein in a given protein complex

Let p_j be a patient

Let q_k be a normal

Let $\Delta_{i,j,k} = Expr(g_i,p_j) - Expr(g_i,q_k)$

Test whether $\Delta_{i,j,k}$ is a distribution with mean 0

Goh & Wong. **Advancing clinical proteomics via analysis based on biological complexes: A tale of five paradigms**. *Journal of Proteome Research*, 15(9):3167--3179, July 2016.

69

Methods to compare with

Network-based methods

- *Hypergeometric enrichment (HE)*
- *Direct group analysis (DG), similar to GSEA*
- *qPSP*
- *PFSNET*
- *ESSNET*

Standard t-test on individual proteins (SP)

70

Simulated data

Simulated datasets from Langley and Mayr

Both D1.2 and D2.2 have 100 small-sized simulated datasets, each with 20% significant features

Effect sizes of these differential features are sampled from one out of five possibilities (20%, 50%, 80%, 100% and 200%), increased in one class and not in the other

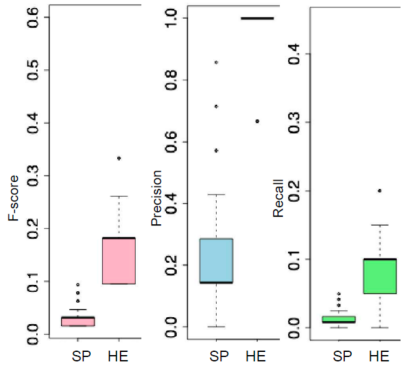
Significant artificial complexes are constructed with various level of purity (i.e. proportion of significant proteins in the complex)

Equal # of non-significant complexes are constructed too

Langley & Mayr, *J. Proteomics*, 129:83-92, 2015

71

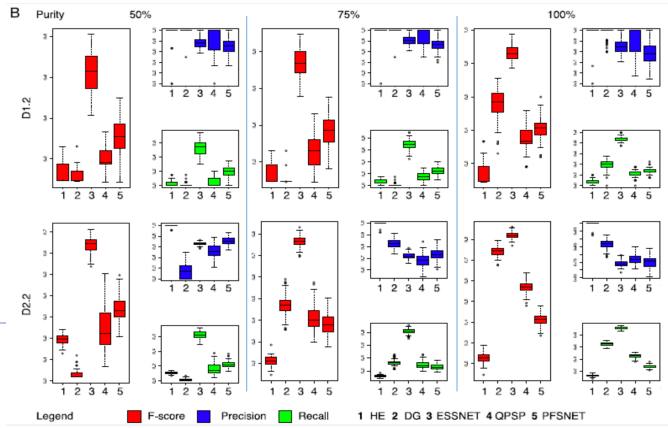
SP shows rather poor performance on simulated data Can network-based methods do better?



Supplementary Figure 1 Single protein (SP) precision-recall performance on D1.2. The f-score (pink), precision (blue) and recall (green) shows that SP performs abysmally on simulated data. HE is shown next to SP as a reference.

72

qPSP / PFSNET / ESSNET perform well on simulated data



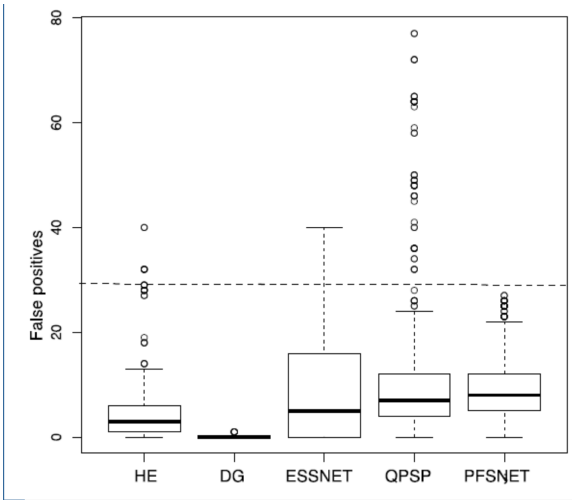
73

Renal cancer control (RCC) data

- 12 samples originating from a human kidney tissue
- 4 biological replicates x 3 technical replicates
- Excellent for evaluating false-positive rates of feature-selection methods
- *Randomly split the 12 runs into two groups*
- *Report of any significant features between the groups must be false positives*

74

All methods control false positives well



Dash line corresponds to expected # of false positives at alpha 0.05 (~30 complexes)

75

Renal cancer data (RC)

- 12 samples are run twice so we have technical replicates over 6 normal and 6 cancer tissues
- Good for testing reproducibility of feature-selection methods
- *A good method should report similar feature sets between replicates*
- Can also test feature-selection stability
- *Apply feature-selection method on subsamples and see whether the same features get selected*

Guo et al. *Nature Medicine*, 21(4):407-413, 2015

76

ESSNET & PFSNET show excellent reproducibility

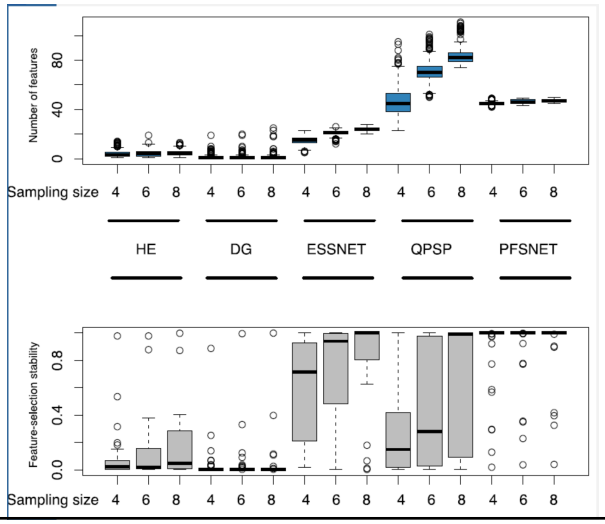
Number of terms	HE	DG	ESSNET	QPSP	PFSNET
Replicate 1	4	1	35	86	45
Replicate 2	6	2	29	75	46
Overlaps	0.25	0.5	0.83	0.66	0.94

HE	DG	ESSNET	QPSP	PFSNET	
1	0.5	0.71	0.86	0.71	HE
	1	1	1	1	DG
		1	0.93	0.98	ESSNET
			1	0.90	QPSP
				1	PFSNET

This table is computed on by applying the methods on the full RC dataset

77

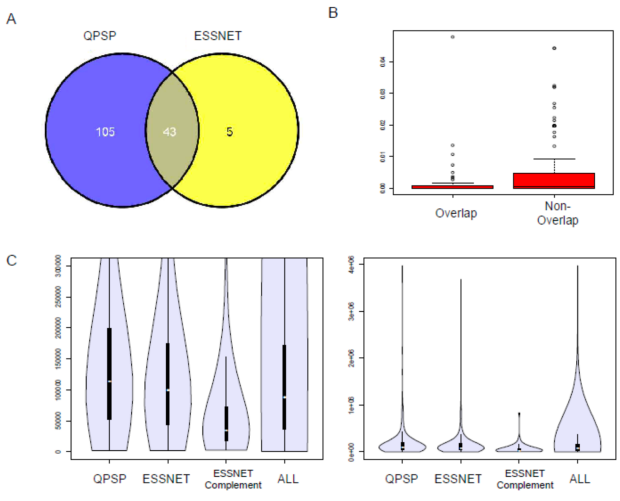
ESSNET & PFSNET show excellent stability



78

ESSNET can assay low-abundance complexes that qPSP cannot

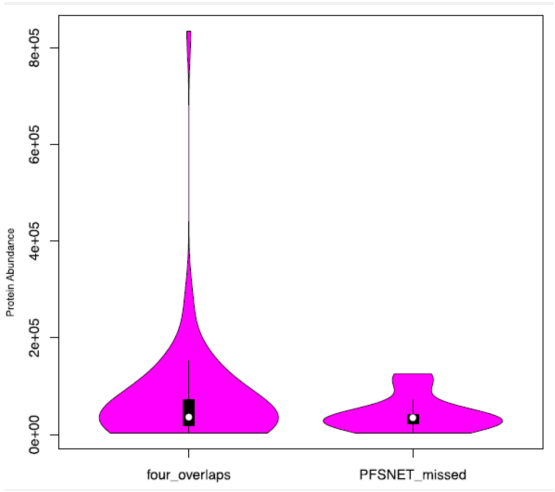
- A: QPSP-ESSNET significant-complex overlaps
- B: P-value distribution for overlapping and non-overlapping QPSP complexes.
- C: Sampling abundance distribution. The left panel is a zoom-in of the right. The y-axis is the protein abundance while the four categories are the distribution of abundances of complexes found in QPSP, ESSNET, ESSNET unique (complement), and all proteins in RC.



79

ESSNET can assay low-abundance complexes that PFSNET cannot

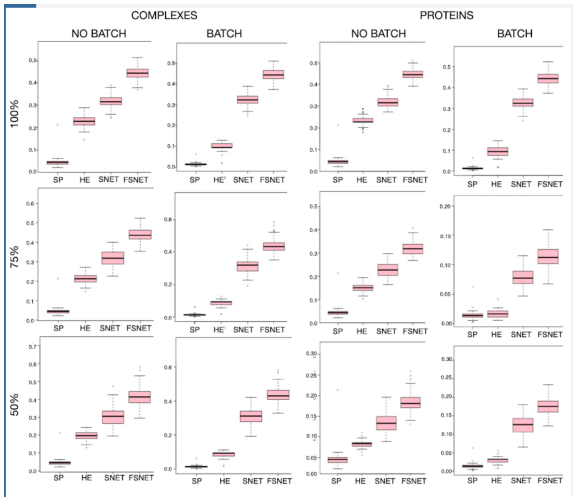
- Of the 5 ESSNET-unique complexes, PFSNET can detect 4; the missed complex consists entirely of low-abundance proteins.
- If p-value threshold is adjusted by Benjamini- Hochberg 5% FDR, PFSNET can detect only 3 of the 5 ESSNET-unique complexes while ESSNET continues to detect them all.



80

Network analysis is robust against batch effects

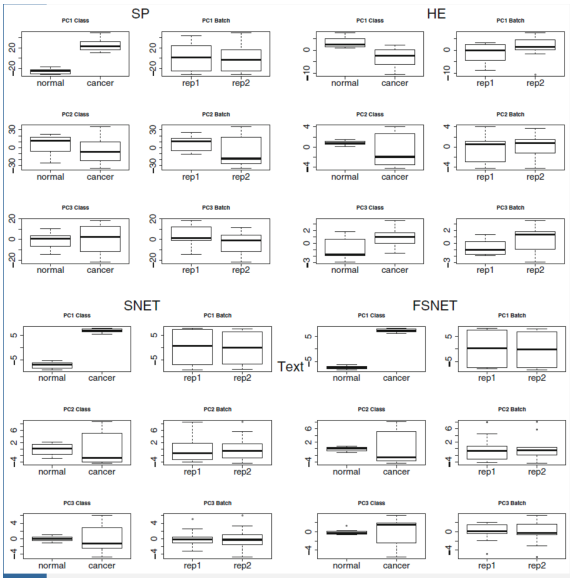
- FSNET is PFSNET using only one-class score; SNET is FSNET with alpha1 = alpha2



Wilson Wen Bin Goh, Limsoon Wong. **Protein complex-based analysis is resistant to the obfuscating consequences of batch effects---a case study in clinical proteomics**. *BMC Genomics*, 18(Suppl 2):142, March 2017.

81

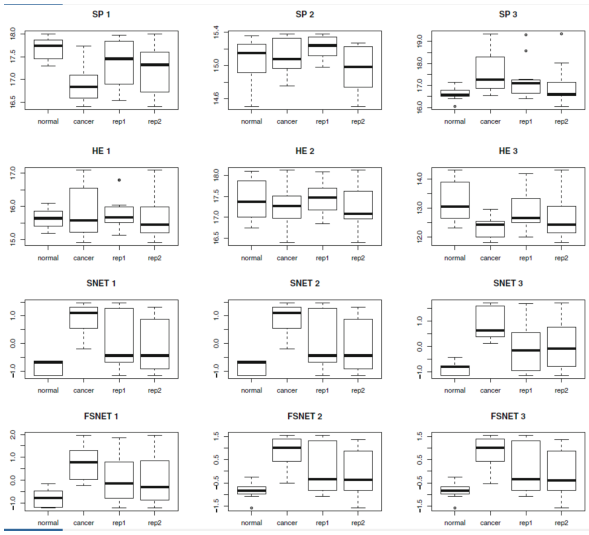
Batch effects "avoid" top PCs produced from protein complexes selected by SNET & FSNET



Wilson Wen Bin Goh, Limsoon Wong. **Protein complex-based analysis is resistant to the obfuscating consequences of batch effects---a case study in clinical proteomics**. *BMC Genomics*, 18(Suppl 2):142, March 2017.

82

Batch effects "avoid" top protein complexes selected by SNET & FSNET

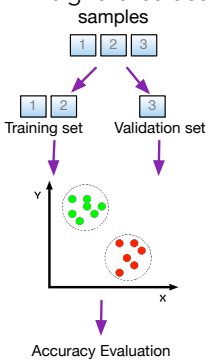


Wilson Wen Bin Goh, Limsoon Wong. **Protein complex-based analysis is resistant to the obfuscating consequences of batch effects---a case study in clinical proteomics**. *BMC Genomics*, 18(Suppl 2):142, March 2017.

83

Networks are very powerful for use with machine learning algorithms

We determine CV accuracy for a signature, and then calculate whether random signatures beat it to get the CV p-val.



Method	Number features	CV accuracy	CV p-val	CV Accuracy/p val
SP	1124	0.98	0.91	1.08
HE	162	0.98	0.91	1.08
SNET	21	0.84	0.06	14.00
FSNET	36	0.96	0.06	16.00
PFSNET	65	0.92	0.06	15.33

Very promising. Network feature score-metrics are very specific. Random features cannot outperform it

Goh & Wong. Evaluating feature-selection stability in next-generation proteomics. *Journal of Bioinformatics and Computational Biology*, 2016

84

Where are we on this now?

We are developing an exciting new missing protein method using networks called PROTREC

Further exploiting the specific use of networks as high-quality input features for ML/AI development

85

Key Takeaway

CONTEXTUALIZATION INTO PROTEIN COMPLEXES CAN HELP WITH CONSISTENCY & REPRODUCIBILITY ISSUES IN PROTEOMIC PROFILE ANALYSIS

IT CAN HELP WITH COVERAGE ISSUE TOO

IT CAN ALSO HELP WITH CONSISTENCY & REPRODUCIBILITY ISSUES IN GENE EXPRESSION PROFILE ANALYSIS

IT CAN MITIGATE AGAINST BATCH EFFECTS

THEY ARE USEFUL FEATURES FOR MACHINE LEARNING

86

(Some) relevant lab publications

- ****Goh WWB**, Zhao Y, Sue ACH, Guo T, Wong LS. Proteomic investigation of intra-tumor heterogeneity using network-based contextualizations, Journal of Proteomics, 30(206):103446, Jul 2019 #
- **Goh WWB**, Wong LS. NetProt: Complex-based feature selection. Journal of Proteome Research, 16(8):3102–3112, June 2017 #
- **Goh WWB**, Wong LS. Integrating networks and proteomics: moving forward. Trends in Biotechnology, 34(12):951-959, Dec 2016 ##
- **Goh WWB**, Wong LS. Design principles for clinical network-based proteomics. Drug Discovery Today, 21(7):1130-1138, Jul 2016 ##
- **Goh WWB**, Wong LS. Advancing clinical proteomics via networks: A tale of five paradigms. Journal of Proteome Research, 15(9):3167-3179, Jul 2016 #

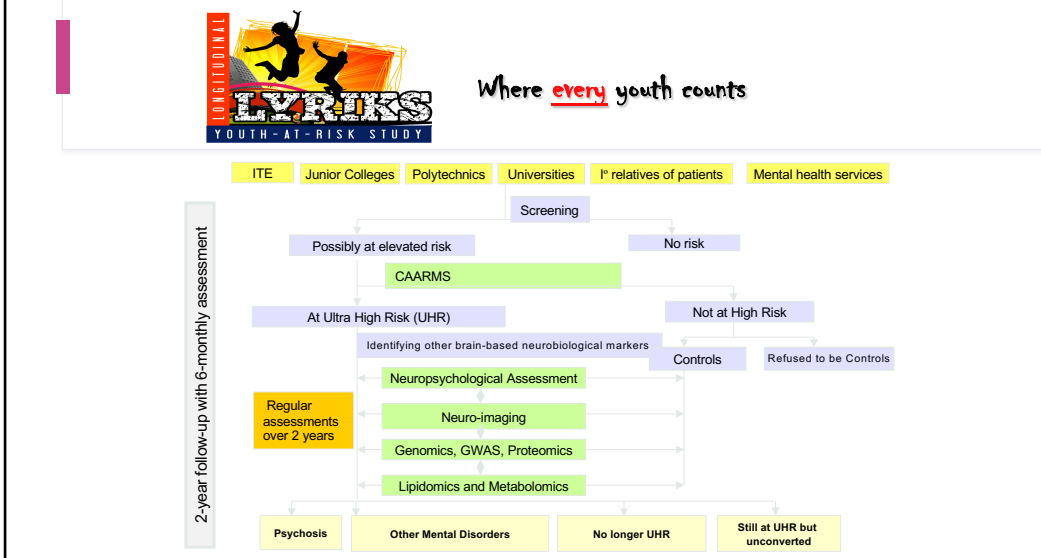
87

Longitudinal study of mental health risk in Singaporean youth

Big Data Integration and AI/ML

Joint work with my PhD student Samuel Tan, Jimmy Lee (IMH/LKC), Limsoon Wong (NUS) and Nikola Kasabov (KEDRI, AUT)

88



89

The LYRIKS advantage (sound of music)

- Largest cohort of UHR subjects from a single site
- Most comprehensive assessments – clinical, imaging, neurocognition and molecular
- Potential to correlate findings across domains
- Relatively free from confounders such as substances and urbanicity

90

The questions I have

- Can ML/AI succeed statistical modelling?
- Amongst next-generation AI models, which dominates?
- Can molecular data augment behavioral data?
- Which combinations of features are powerful?
- Can we mine the ML/AI for mechanistic insights?

91

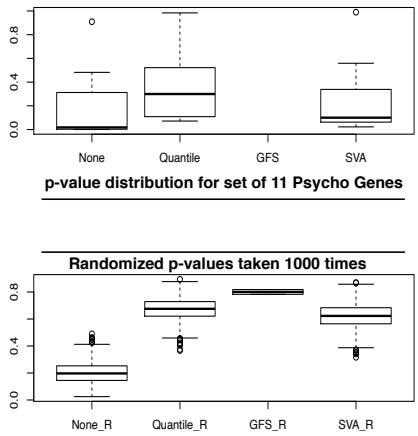
The team (right now…)

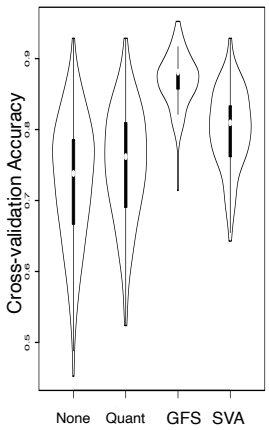
Focus	Investigator	Location
Data Science (Lead)	Wilson Goh	SBS
Clinical	Jimmy Lee	IMH/LKC
AI (SNNs)	Nikola Kasabov Maryam Doborjeh	KEDRI/AUT
RNA-Seq	Foo Jia Nee	GIS/LKC
Metabonomics	Wang Yulan	SPC/LKC
Proteomics	Tiannan Guo	Guomics/Westlake U
Bioinformatics	Limsoon Wong	SOC/NUS

92

Fuzzy scoring (GFS) boosts prediction reproducibility

Recall that GFS is step 1 of PFSNET





Goh et al. Can peripheral blood-derived gene expressions genetically characterize high risk subjects for psychosis? Computational Psychiatry 2017

93

Promising early results using networks

- PFSNET features + random forest --- evaluated using leave-one-out cross-validation
- F-mean considers precision and recall
- Youden's J is (sensitivity + specificity) - 1

	Complex as features		Genes as features				
	PFSnet (ModComplexScore)	**PFSnet (WeightedDelta)**	WTT (Avg Intensity)	PFSnet (Avg Intensity)	WTT (Avg Intensity)	Random400 (Avg Intensity)	Random1300 (Avg Intensity)
True Positive	43	**50**	47	43	46	43	44
False Positive	10	**8**	10	8	10	8	8
True Negative	18	**20**	18	20	20	20	20
False Negative	13	**6**	9	13	10	13	12
F measure	0.79	**0.88**	0.81	0.80	0.81	0.80	0.81
Youden's J	0.41	**0.61**	0.48	0.48	0.49	0.48	0.50

94

Where are we on this now?

BEHAVIORAL DATA ANALYSIS FIRST (TEST PREDICTABILITY AND EARLIEST DETECTABILITY WINDOW)

DEMONSTRATE THAT NETWORKS WORK WELL ON MOLECULAR INFORMATION E.G. GENOMICS AND PROTEOMICS

RECRUITMENT! LOOKING FOR 2 RESEARCH ASSOCIATES AND 1 RESEARCH ASSISTANT TO JOIN THIS PROJECT!!! ADDITIONAL PHD STUDENTS IN AUT ALSO WANTED!

95

(Some) relevant lab publications

- **Ho. SY, Phua K, Wong L, **Goh WWB**. Extensions of the external validation for checking learned model interpretability and generalizability, Patterns (Cell Press), Accepted
- **Ho SY, Sze CC, Wong L, **Goh WWB**. What can Venn diagrams teach us about doing data science better? International Journal of Data Science and Analytics, Accepted
- **Ho SY, Wong L, **Goh WWB**. Avoid oversimplifications in machine learning: Going beyond the class-prediction accuracy. Patterns (Cell Press), 1(2):100025, May 2020
- Goh WWB, Sng J, Yee JY, See YM, Lee TS, Wong LS, Lee J. Can peripheral blood-derived gene expressions genetically characterize high risk subjects for psychosis? Computational Psychiatry, 0(1):1-16, Oct 2017

96

Our other project areas..

Data science (TCM)	AI in Education	Statistics	Meta-ML problems
• Predicting patient disease using TCM features • Feature engineering and database design	• Performance prediction • Intelligent coaching platforms • Co-evolution with pedagogy • Graph literacy • Bio-data science education	• Anna Karenina Principle • Relooking Neyman-Pearson approaches • Stability and reproducibility problems	• Sloppy models • No-free lunch theorem • Rashomon set problems • Doppelganger effect • Better evaluation metrics • Better feature engineering

97

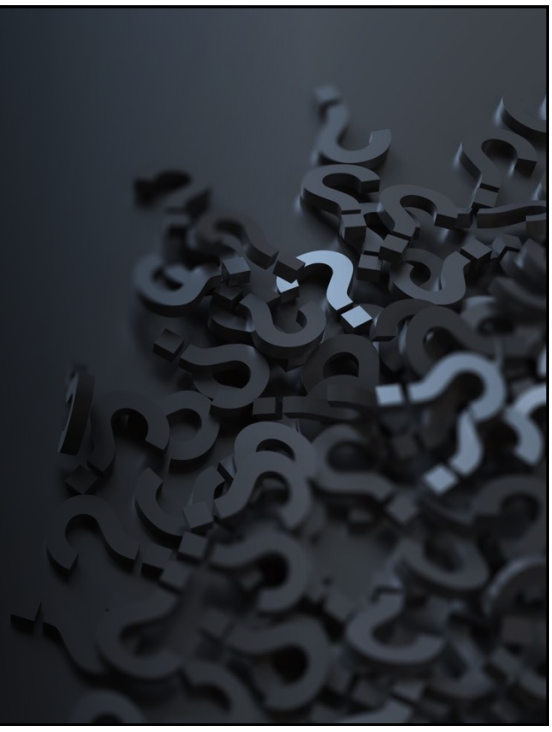
Acknowledgements

- MOE ACRF 1 and 2
- National Research Foundation, Singapore
- Singapore Data Science Consortium
- National Medical Research Council
- My supportive colleagues in SBS, LKC and NUS
- My amazing team!

98

Questions?

Thanks for listening



99