Statistical analysis for NAPPA

Yunro Chung, Ph.D.
Assistant Professor
Arizona State University

Schedule

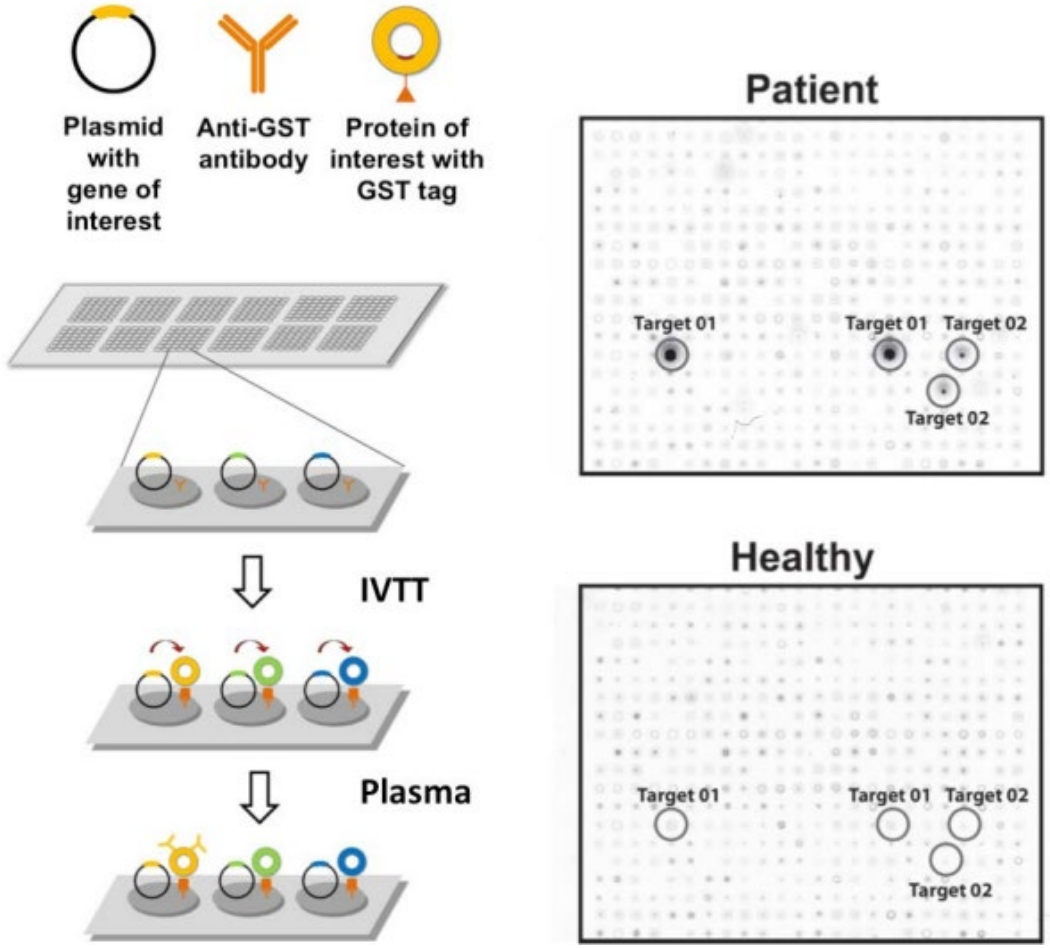
Date	Time	Topic
5/19/2025	10-10:30am	L1: Overview & Normalization
	10:30-12pm	L2: Evaluation of biomarkers
	12-1pm	Lunch
	1-3 pm	R practice for L1 and L2
	3-4 pm	Q&A
5/20/2025	10-11am	L3: Statistical hypothesis testing
	11-12pm	L4: Machine learning
	12-1pm	Lunch Break
	1-3 pm	R practice for L3 and L4
	3-4 pm	Q&A

- The training will be for both non-data and data specialists.
- The afternoon sessions will be primary for data specialists, but non-data specialists are also welcome to join.

Overview: NAPPA

https://www.youtube.com/watch?v=mgi408B6Cmg

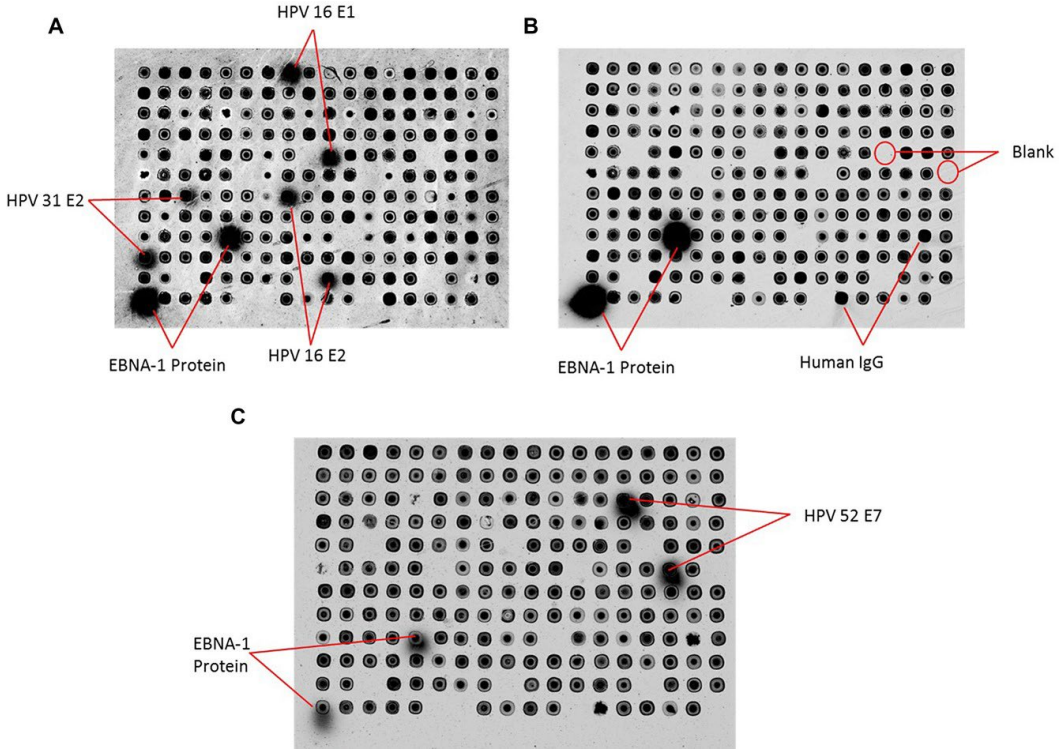
Standard NAPPA

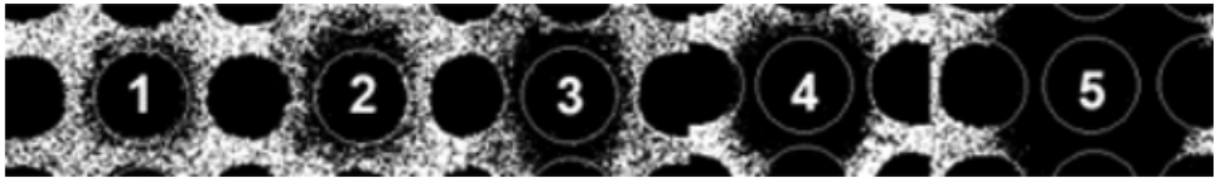


- Glass slide surface
- Scan each slide to create image files & measure each spot intensity of these files.

Wang, J., 2015. *Cancer autoantibody biomarker discovery and validation using nucleic acid programmable protein array* (Doctoral dissertation, Arizona State University).

Standard NAPPA - Halo effect



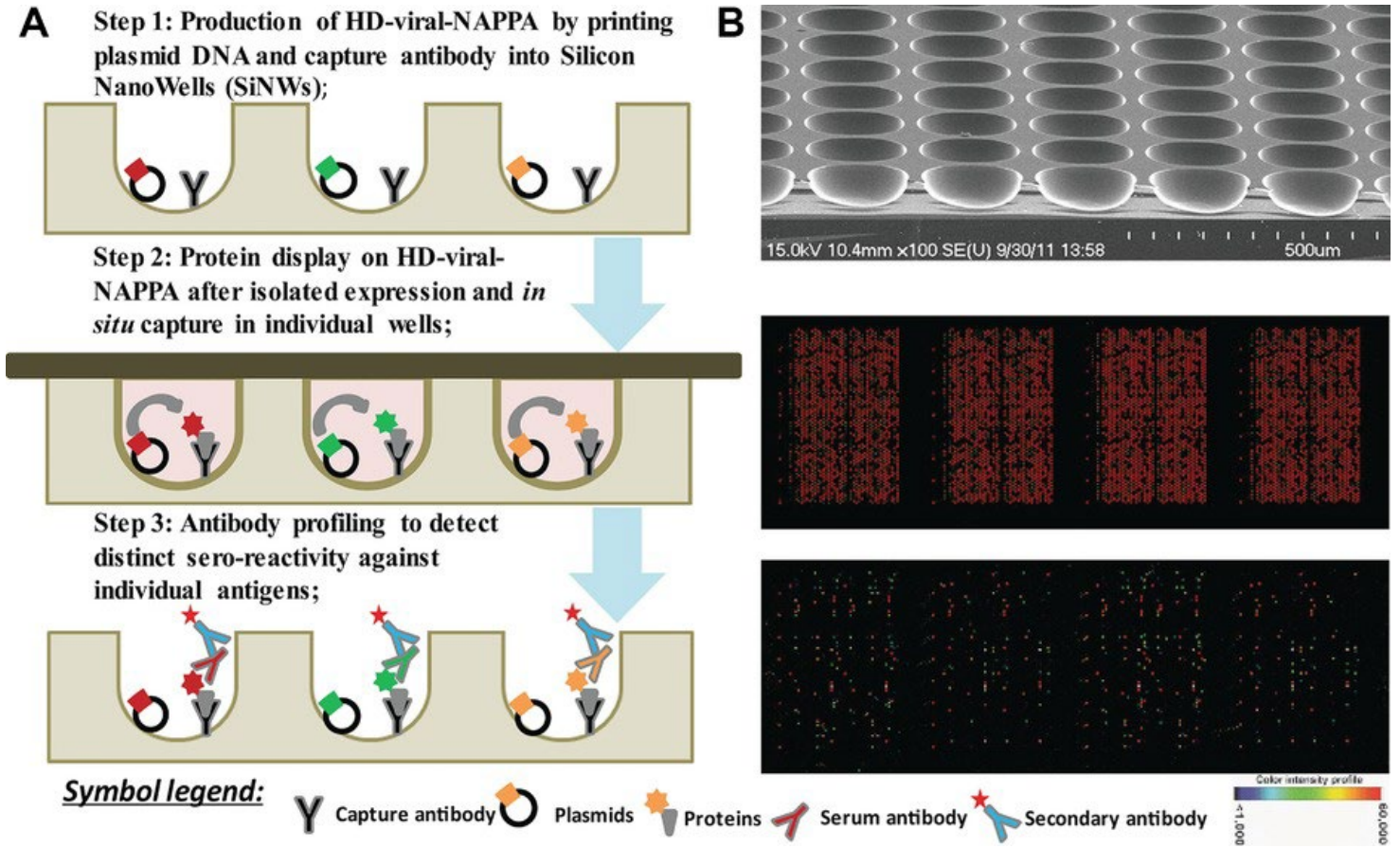


- Before data analysis, visual inspections are recommended, and proteins near positive controls may be removed.

Ewaisha, R., Meshay, I., Resnik, J., Katchman, B.A. and Anderson, K.S., 2016. Programmable protein arrays for immunoprofiling HPV-associated cancers. *Proteomics*, *16*(8), pp.1215-1224.

Rivera, R., Wang, J., Yu, X., Demirkan, G., Hopper, M., Bian, X., Tahsin, T., Magee, D.M., Qiu, J., LaBaer, J. and Wallstrom, G., 2017. Automatic identification and quantification of extra-well fluorescence in microarray images. *Journal of Proteome Research*, *16*(11), pp.3969-3977.

High-Density (HD) NAPPA



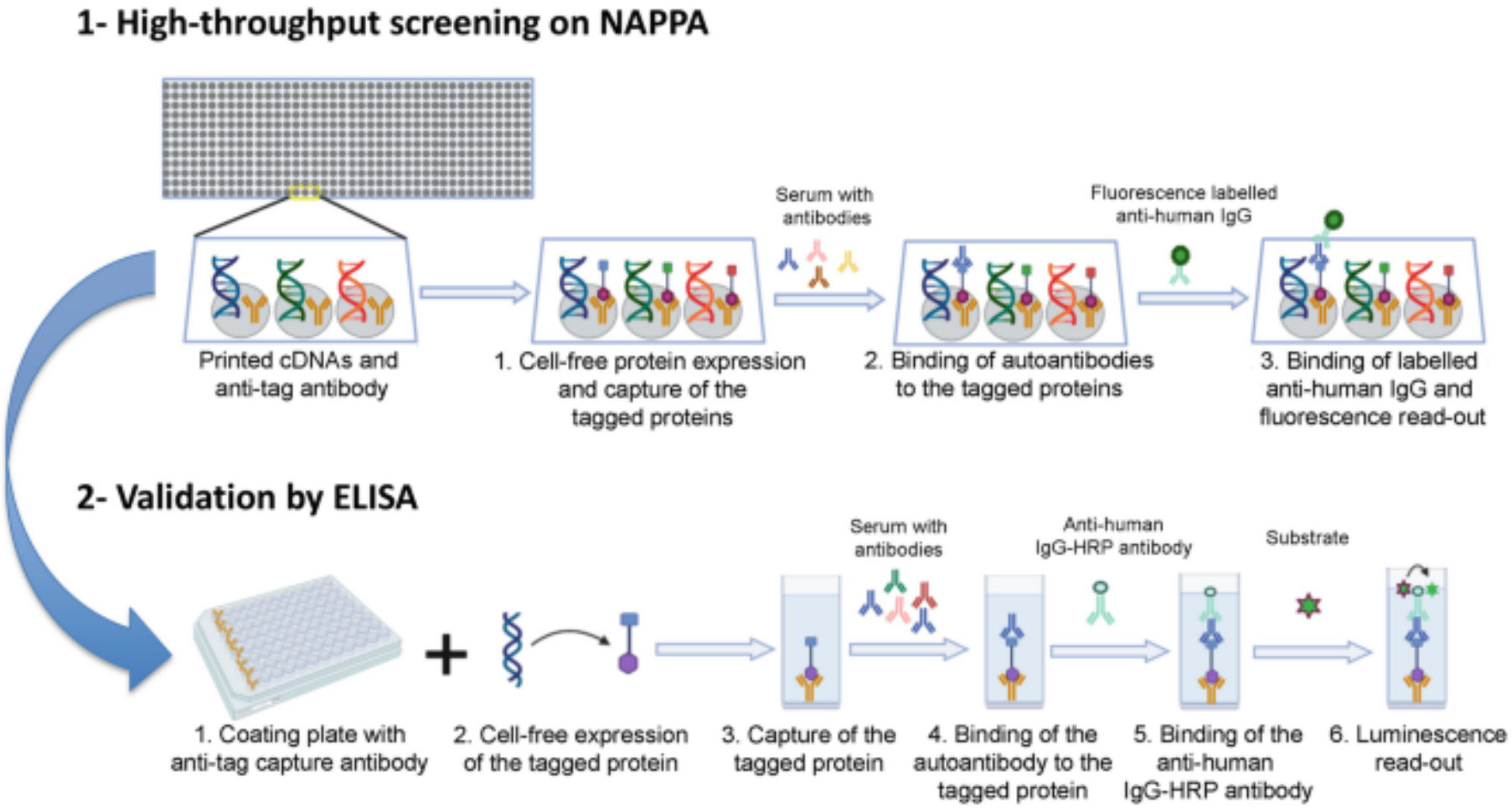
- Nano well technology to prevent the halo effect.
- Still visual inspections are recommended

Bian, X., Wiktor, P., Kahn, P., Brunner, A., Khela, A., Karthikeyan, K., Barker, K., Yu, X., Magee, M., Wasserfall, C.H. and Gibson, D., 2015. Antiviral antibody profiling by high-density protein arrays. *Proteomics*, *15*(12), pp.2136-2145.

Study design

- Case-control studies
- Goal is to identify novel biomarkers for disease screening / diagnosis.
  - More generally, any binary outcome D=1 versus D=0.
- Ideally, age, gender, other risk factors is matched between the two groups.
  - Example of poor study designs: cases are significantly younger than controls; cases are smokers, whereas controls are non-smokers, etc.
- Randomize NAPPA experiments.
  - Example of poor study designs: day 1 is for all case samples; day 2 is for all control samples.
- ***A careful study design is essential.***

Discover potentially useful biomarkers



Sibani, S. and LaBaer, J., 2011. Immunoprofiling using NAPPA protein microarrays. *Protein Microarray for Disease Analysis: Methods and Protocols*, pp.149-161.

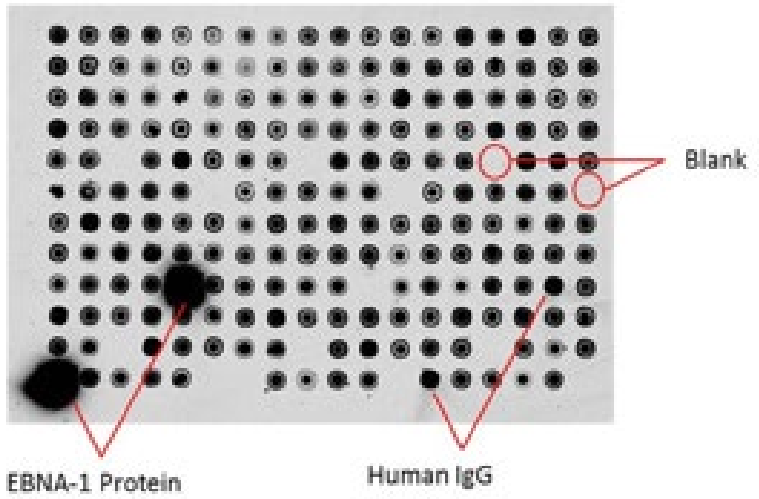
Data Analysis

- The data structure is similar between NAPPA and microarray studies.
- Many statistical and computational tools were developed for analyzing microarray data, particularly in the 2000s.
- These tools may or may not be used for NAPPA data.
- Caution: NAPPA data is more skewed than microarray data.

Normalizations

Normalizations

- NAPPA usually includes control spots:
  - Positive control spots
  - Negative control spots
  - Empty spots



- To reduce biases from technical variations, we have used the quantile or median normalization methods, particularly using these control spots.
- Suppose there are 3 subjects with 5 proteins, where P3 and P4 are positive control spots, and P5 and P6 are empty spots.

Sample	P1	P2	P3	P4	P5	P6
1	10	10	10	20	1	3
2	20	30	30	40	2	1
3	30	40	60	30	3	4

Quantile Normalization

- Normalized intensity$_{ij}$ = $(X_{ij}-BC_i)/(MED-BC_i)$ (ith sample, jth protein)
  - MED=media of all proteins except control spots.

    Median of (10,20,30,10,30,40)=25
  - BC_i=first quartile of ith sample's empty spots.

Sample	P1	P2	P3	P4	P5	P6	BC of P5 & P6	Normalized P1	Normalized P2
1	10	10	10	20	1	3	1.5	(10-1.5)/(25-1.5) =0.362	(10-1.5)/(25-1.5) =0.362
2	20	30	30	40	2	1	1.25	(20-1.25)/(25-1.25) =0.789	(30-1.25)/(25-1.25) =1.211
3	30	40	60	30	3	4	3.25	(30-3.25)/(25-3.25) =1.230	(40-3.25)/(25-3.25) =1.690

Meidan Normalization

- Normalized intensity$_{ij}$= X_{ij}/(MED^{+}_{i}- MED^{-}_{i}) (ith sample, jth protein)
  - MED^{+}_{i}=median of ith sample's positive control spots.
  - MED^{-}_{i}=median of ith sample's empty spots.

Sample	P1	P2	P3	P4	MED^{+} of P3 & P4	P5	P6	MED^{-} of P5 & P6	Normalized P1	Normalized P2
1	10	10	10	20	15	1	3	2	10/(15-2) =0.769	10/(15-3) =0.769
2	20	30	30	40	35	2	1	1.5	20/(35-1.5) =0.597	30/(35-1.5) =0.896
3	30	40	60	30	45	3	4	3.5	30/(45-3.5) =0.723	40/(45-3.5) =0.964

Summary

- After these normalizations, $\log_2(X)$ can be used.
- Other normalization methods can also be used.
- What is the best normalization method?
  - There is no universal approach that consistently outperforms the others.
- One alternative:
  1. After conducting a NAPPA experiment, use a particular normalization method and select some biomarkers.
  2. Validate them using ELISA.
  3. If this normalization performs well, the two results are aligned (but may not be perfect due to a chance of false discovery).

  An optimizing process may be needed to select the best normalization method.

Dataset https://github.com/yunro-chung/NAPPA

- 45 breast cancer and 45 control samples.
- There are 84 rows x 28 columns = 2352 spots:
  - Majorities are spots for proteins of interests.
  - Few of them are positive/negative control spots,
    - 25 positive control spots: 2 EBNA & 23 hIgG
    - 29 empty spots: 2 MM & 23 NON-SPOT
    - Other control spots: REG
- Multiple assays can also be used for the same sample if the number of proteins of interests is large, e.g.,
  - assay 1 for proteins 1-2000; assay 2 for proteins 2001-4000, etc.