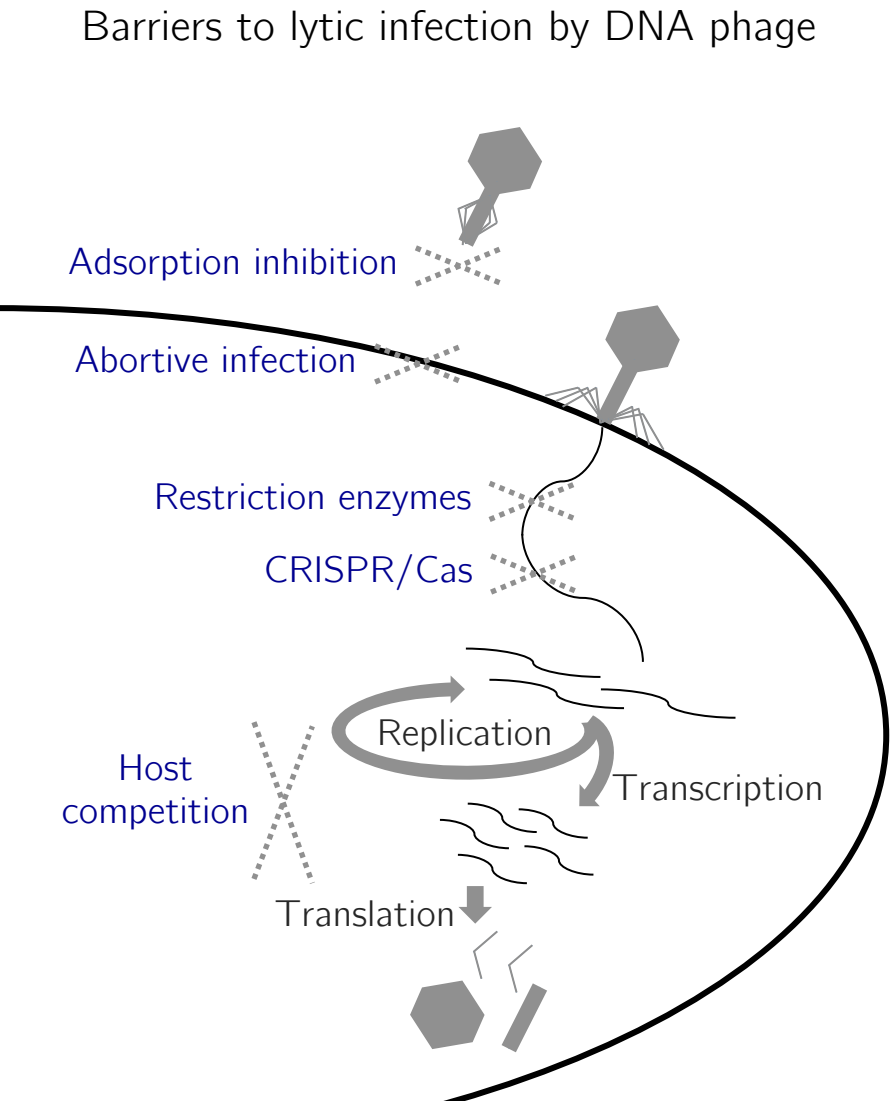
Statistically identifying mechanisms of phage host interactions in the Nahant Collection

Joy Yang
Polz Lab, MIT
CSGF Review – July 13, 2018

Tiny organisms in large numbers have a large impact on our ecosystem

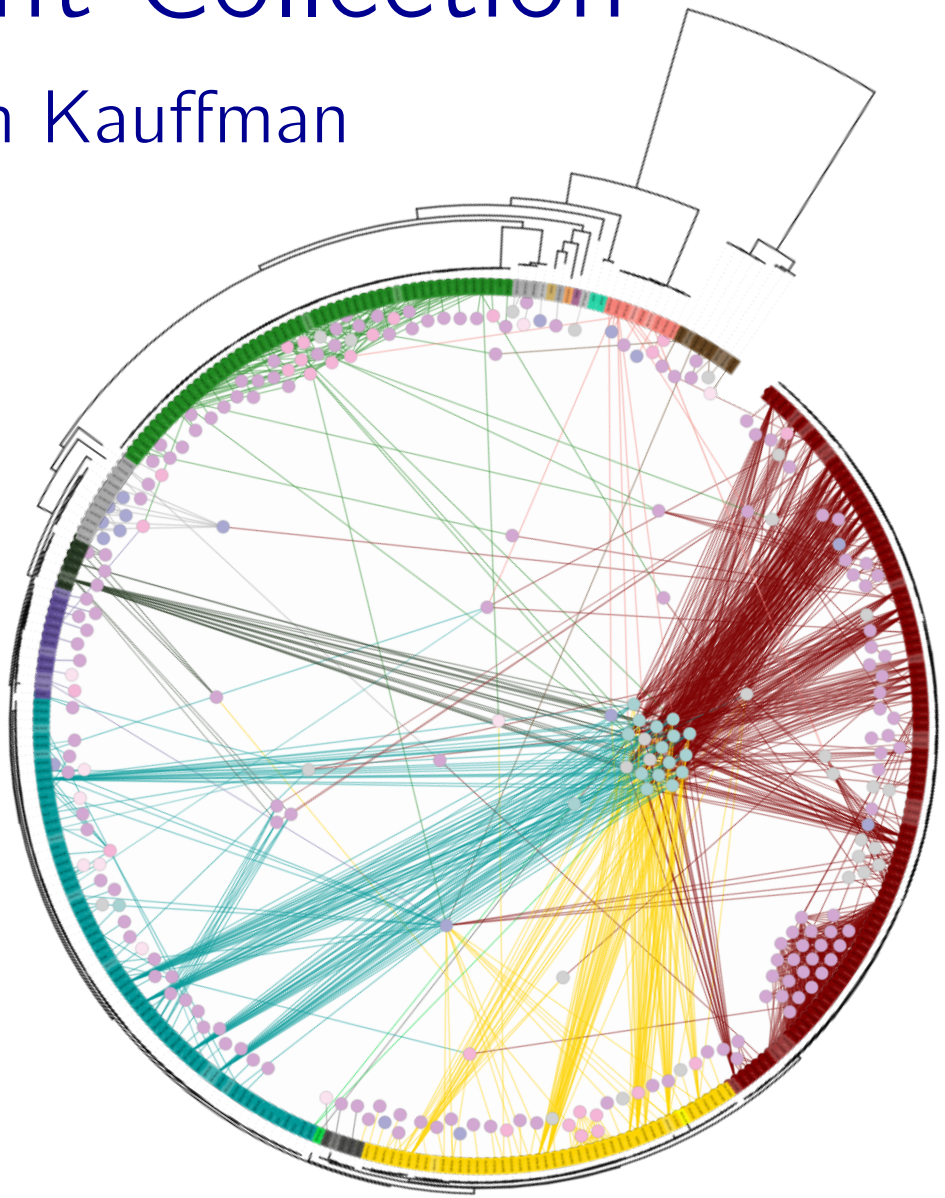


- In 1 mL of sea water
  - 10^6 bacteria
  - 10^7 phage
- 36 million km3 of water in the top 100m of the sea
- Rare events for a single cell do not equate to rare events for the whole population
- The evolutionary arms race rapidly evolves arsenals of largely unexplored mechanisms

The Nahant Collection

Kathryn Kauffman

- Largest phylogenetically resolved phage-host cross-test
- 241 diverse phage
  - nontailed (Tecti-)
  - tailed (Podo-, Myo-, Sipho-)
- 243 hosts, ecologically differentiated
  - E. norv are free-living
  - V. cyc, large-particle specialists
  - V. tas are generalists.
- 1000 phage gene clusters
- 10,000 host gene clusters

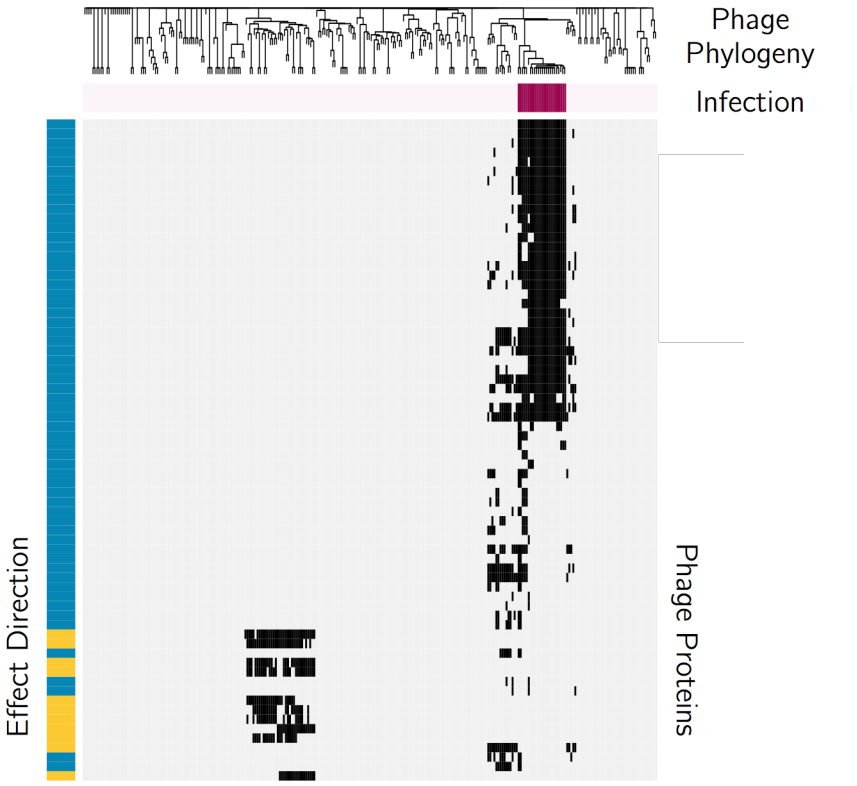


Thesis Aims

- Aim 1: Identify novel mechanisms and infection and defense in the coastal ocean
- Aim 2: Elucidate the role of phage 2.275.O. tRNA during the infection cycle
- Aim 3: Develop curricula for engineering/ statistics outreach

Aim 1: Identifying novel mechanisms and infection and defense in the coastal ocean

Correcting for population structure is important for sifting out relevant signals from spurious correlations



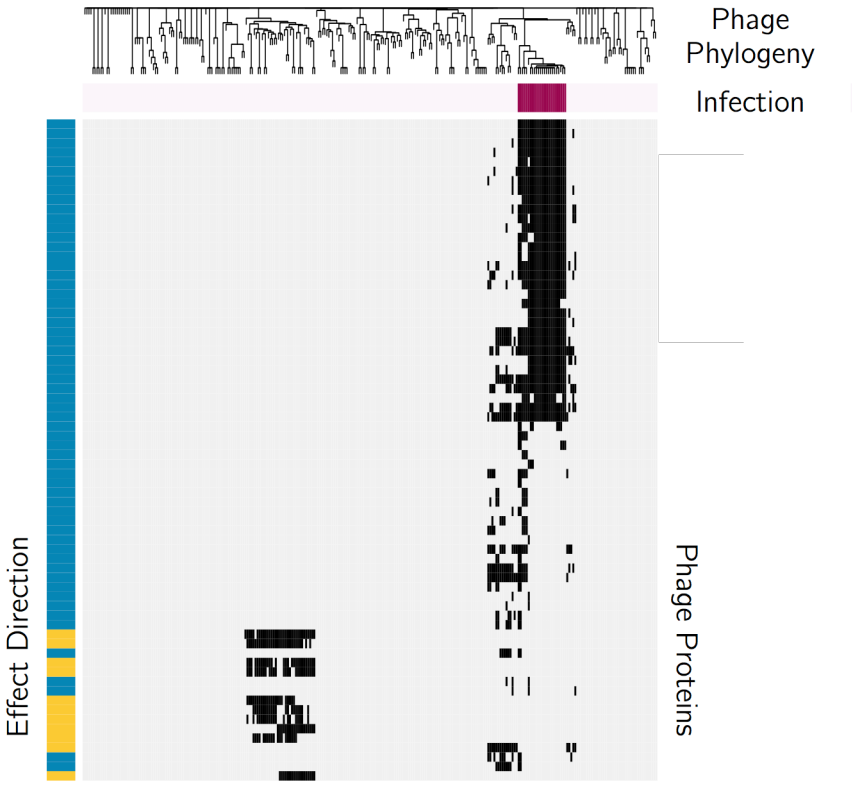
Ordinary Least Squares

Effect Direction
positive
negative

Phage/Host
infects
does not infect

Phage Protein
present
absent

Correcting for population structure is important for sifting out relevant signals from spurious correlations



Ordinary Least Squares

$$Y = X\beta + \varepsilon$$
$$cov(\varepsilon) = \Sigma$$

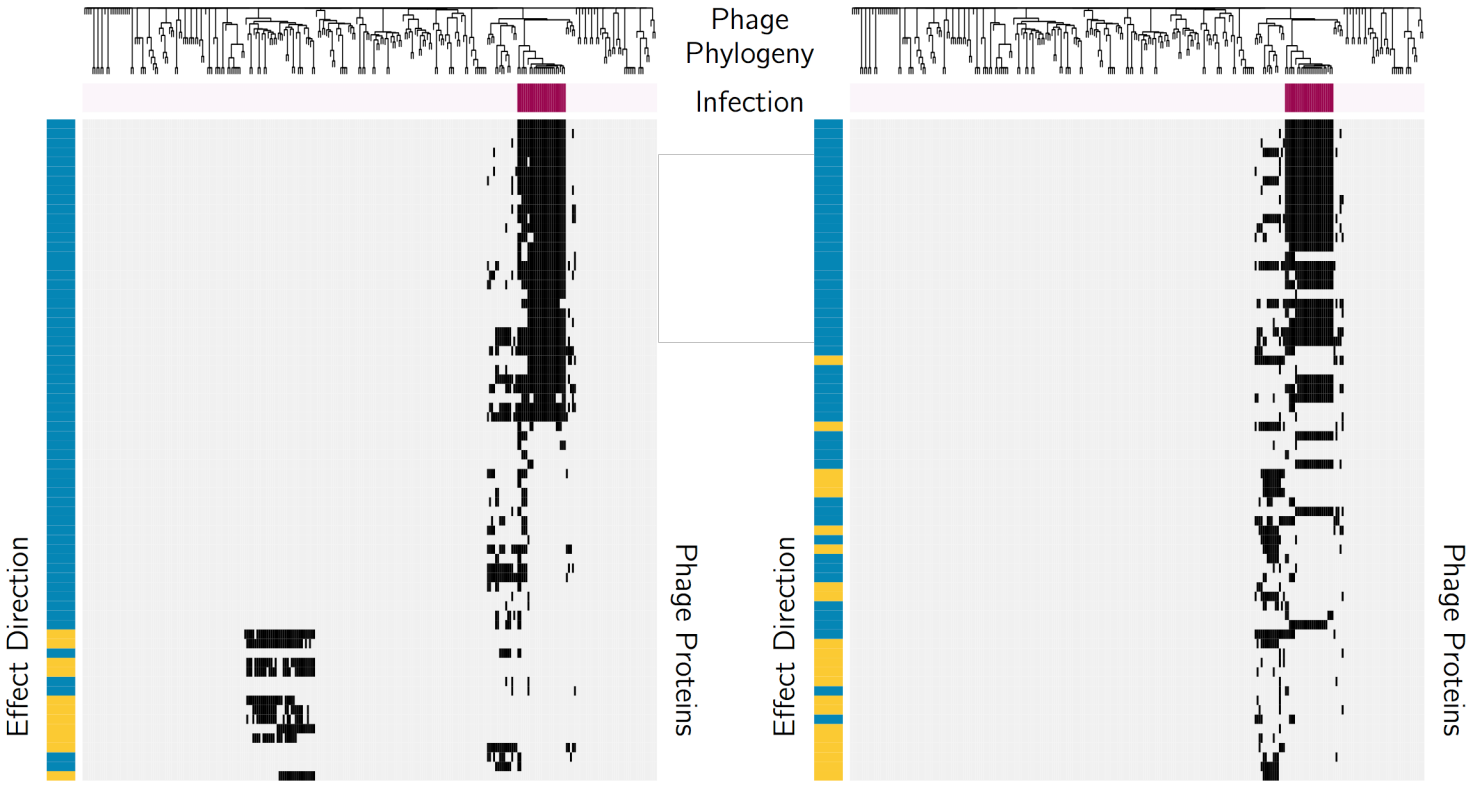
$$\widetilde{Y} = \widetilde{X}\beta + \widetilde{\varepsilon}$$
$$\Sigma^{-1/2}Y = \Sigma^{-1/2}X\beta + \Sigma^{-1/2}\varepsilon$$
$$cov(\widetilde{\varepsilon}) = \Sigma^{-\frac{1}{2}} cov(\varepsilon) \Sigma^{-\frac{1}{2}\prime}$$
$$= \Sigma^{-\frac{1}{2}} \Sigma \Sigma^{-\frac{1}{2}\prime}$$
$$= I$$

Effect Direction
positive
negative

Phage/Host
infects
does not infect

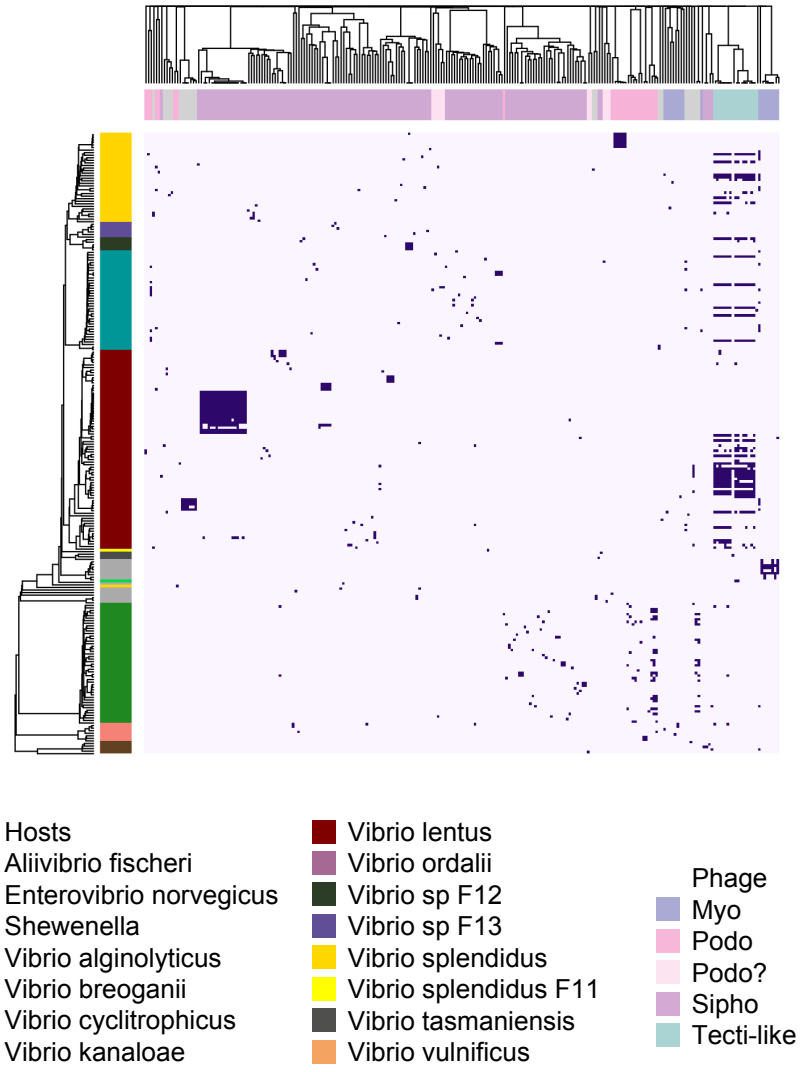
Phage Protein
present
absent

Correcting for population structure is important for sifting out relevant signals from spurious correlations



Generalizing the analysis to the 2D matrix allows us to model interacting phage/host systems

$$Y \sim Bern(\mu)$$
$$logit(vec(\mu)) = \beta_0 + X_v\beta_v + X_h\beta_h + X_{vh}\beta_{vh}$$

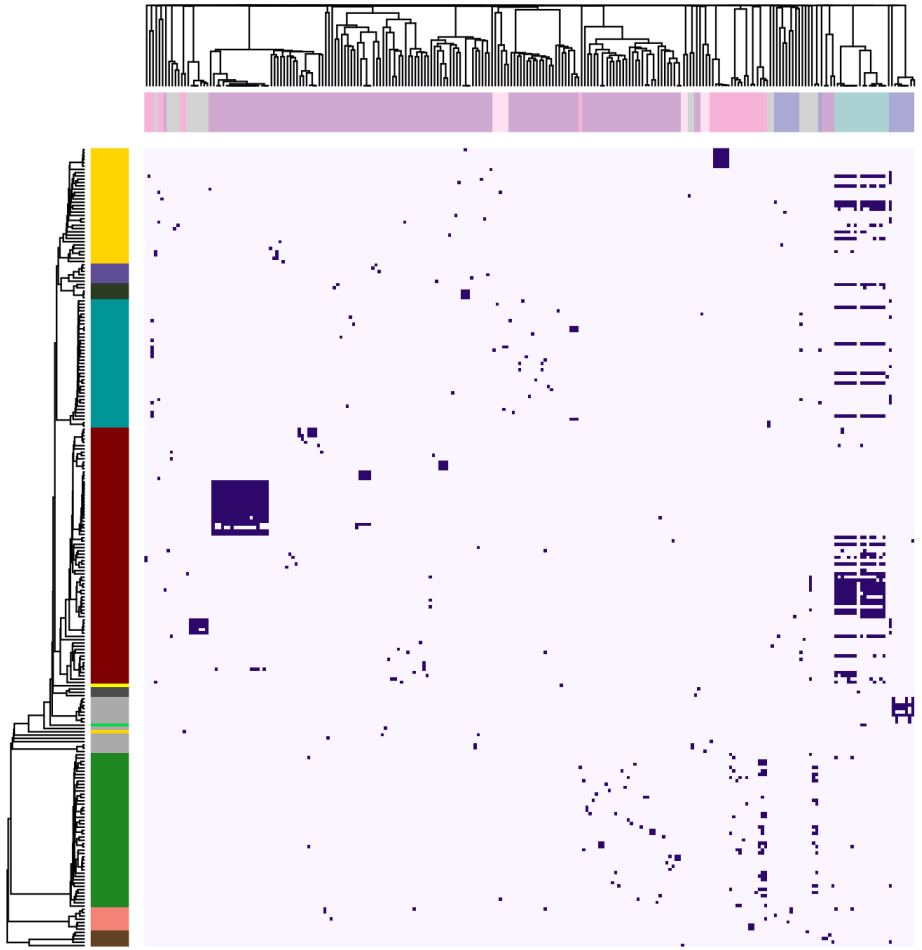


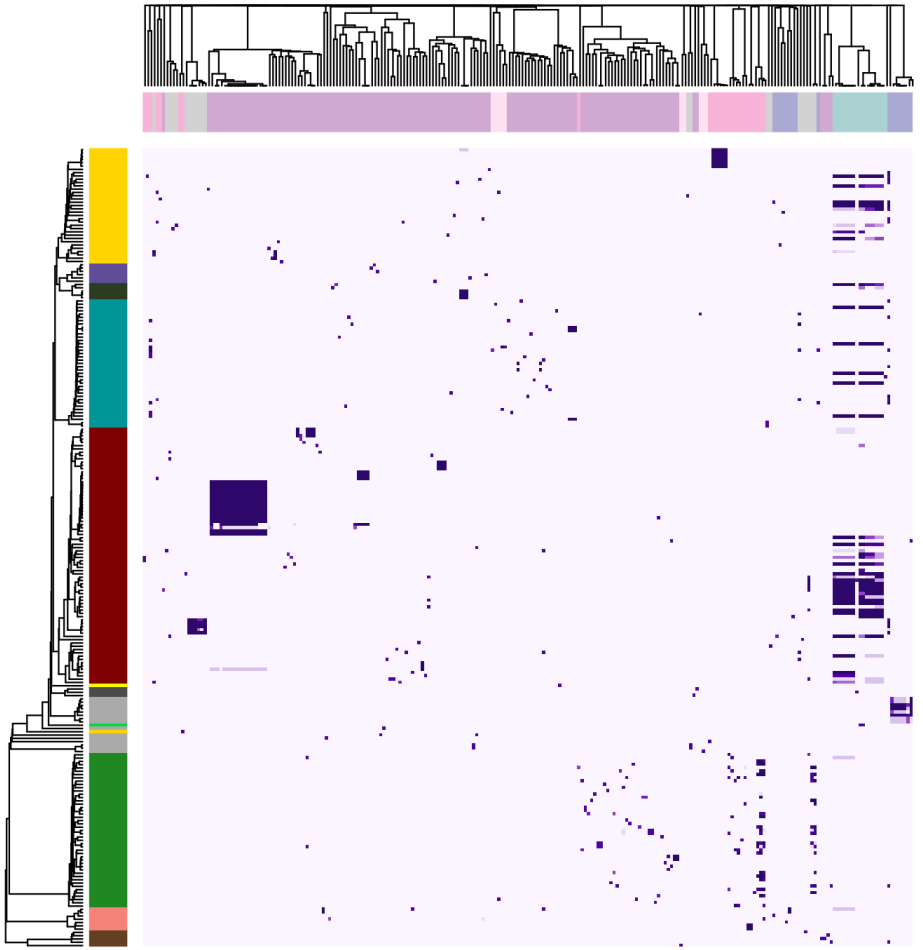
10,000,000 interaction terms

$$
\begin{aligned}
Y &\sim Bern(\mu) \\
logit(vec(\mu)) &= \beta_0 \\
&\quad + X_v\beta_v + X_h\beta_h \\
&\quad + X_{vh}\beta_{vh}
\end{aligned}
$$

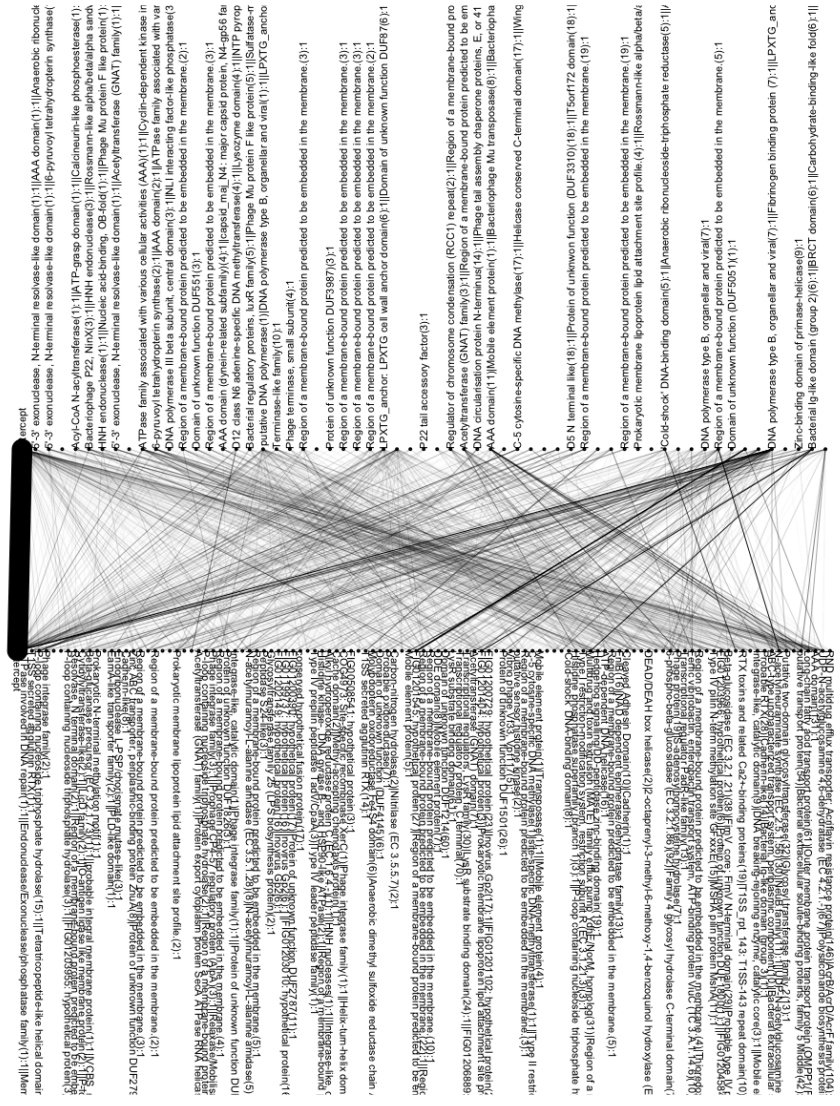
- 1000 virus proteins
- 10,000 host proteins
- 10,000,000 interaction terms
- Y is 243 x 241 ~ 58,000
- To calculate M, predictors ~58,000 x 10,000,000
- ~4.7 TB
- Sparse encoding is ~4 GB
- How do we interpret 10,000,000 coefficients?

Generalizing the analysis to the 2D matrix allows us to model interacting phage/host systems





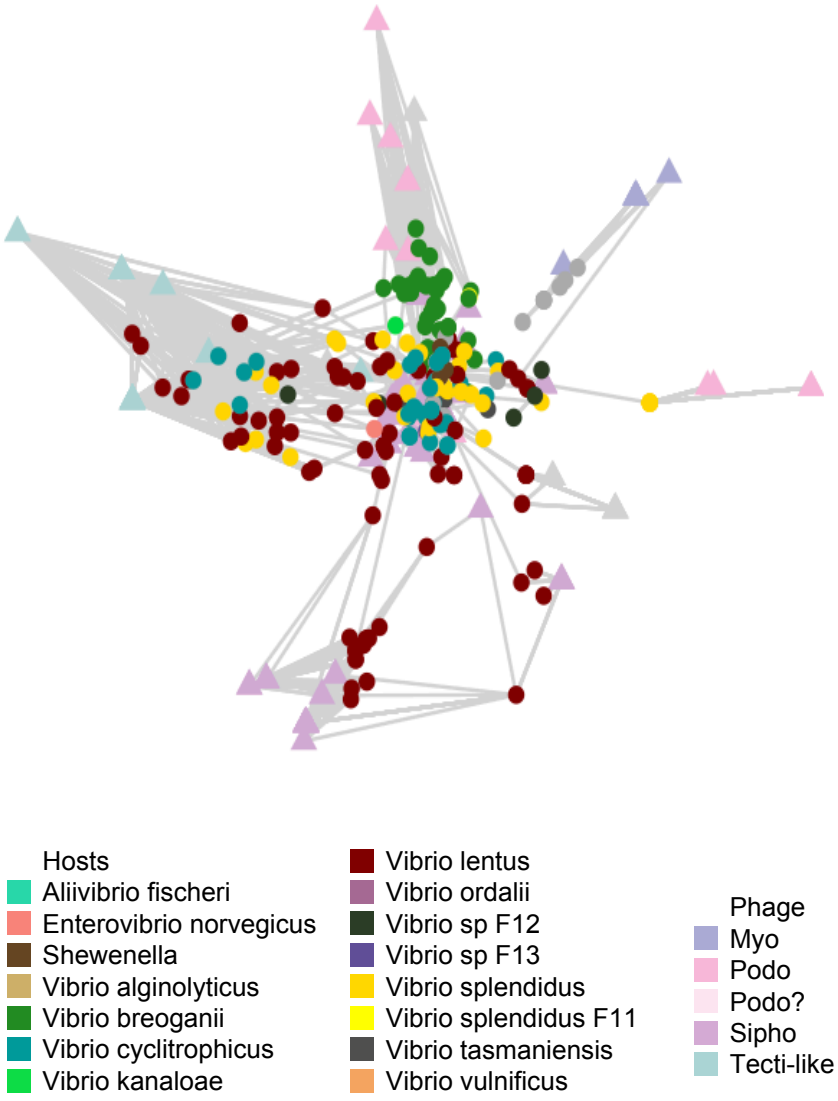
But how do we interpret the coefficients?



Geometric interpretation of projecting phage and hosts into the same space

$$Y \sim Bern(\mu)$$

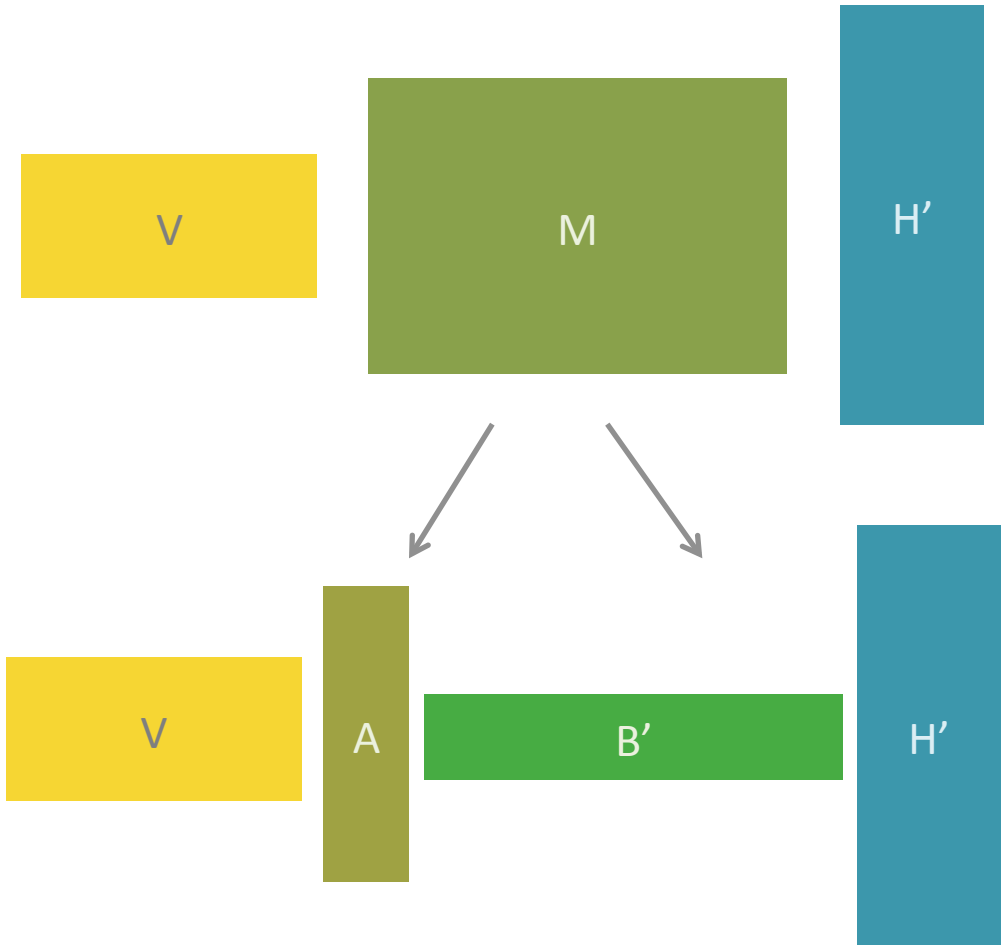
$$logit(\mu) = < A'V', B'H' >$$



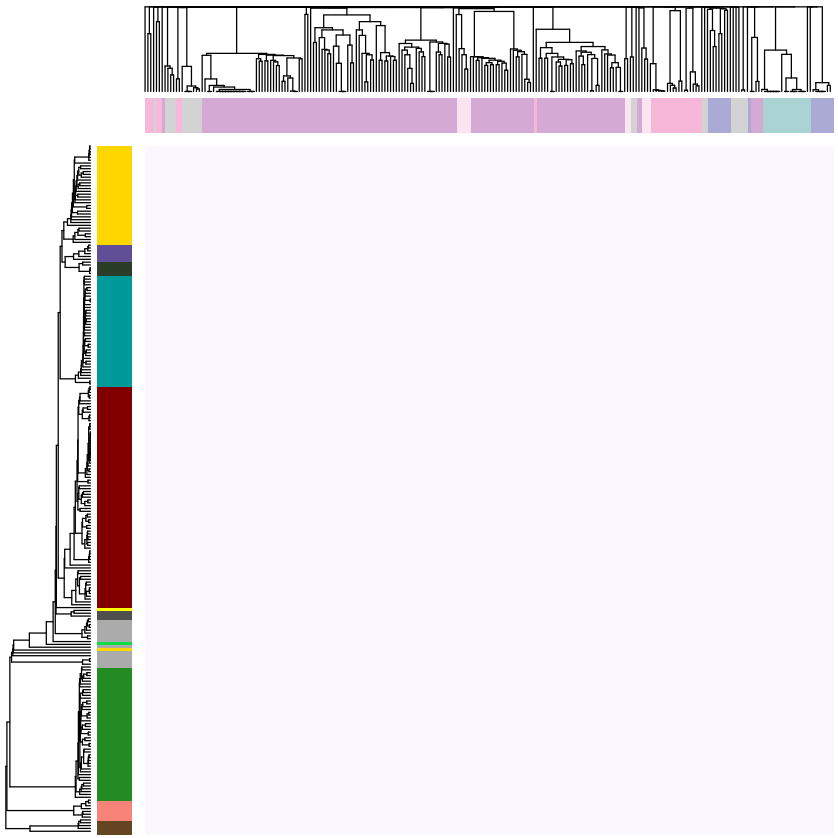
Idea from Philippe Rigollet

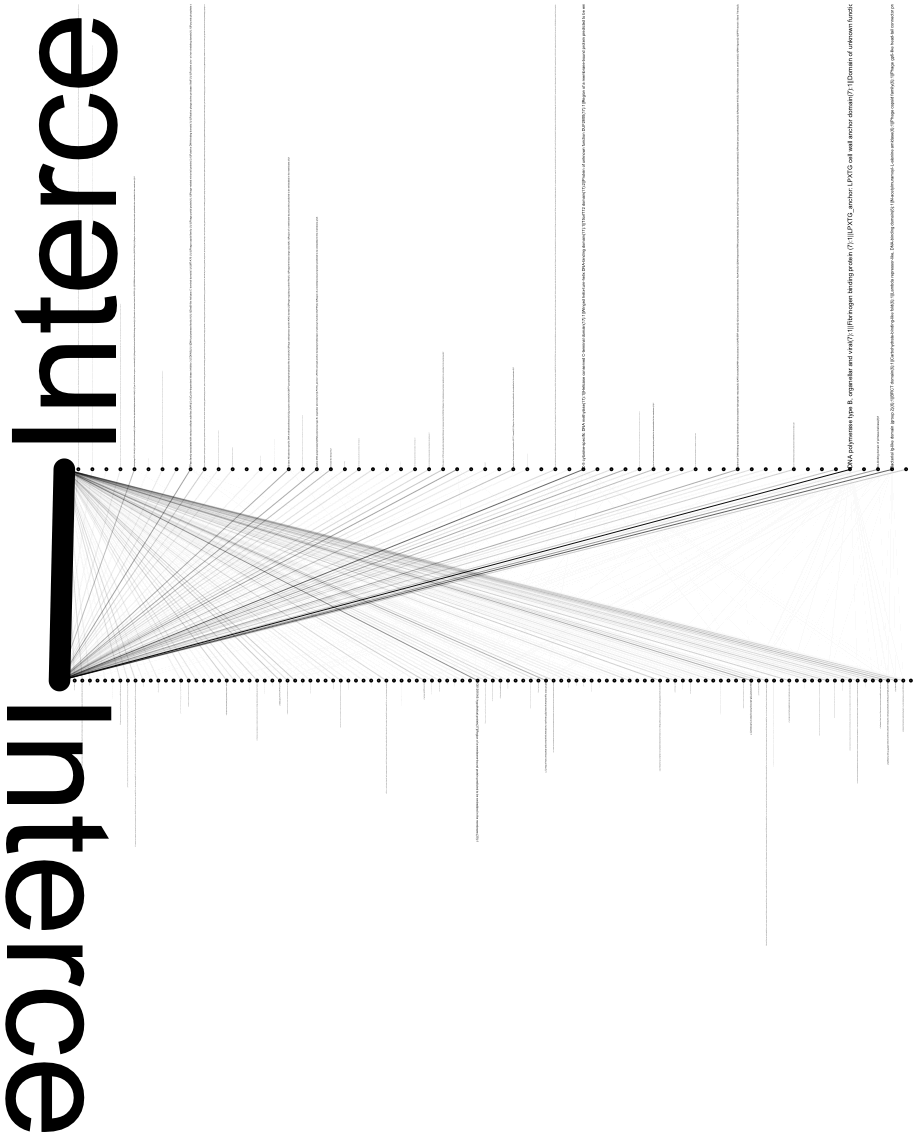
Idea: interpret orthogonal components?

$$
\begin{aligned}
Y &\sim Bern(\mu) \\
logit(vec(\mu)) &= \beta_0 \\
&\quad + X_v\beta_v + X_h\beta_h \\
&\quad + X_{vh}\beta_{vh} \\
logit(\mu) &= VMH' \\
logit(\mu) &= VAB'H' \\
logit(\mu) &= < A'V', B'H' >
\end{aligned}
$$

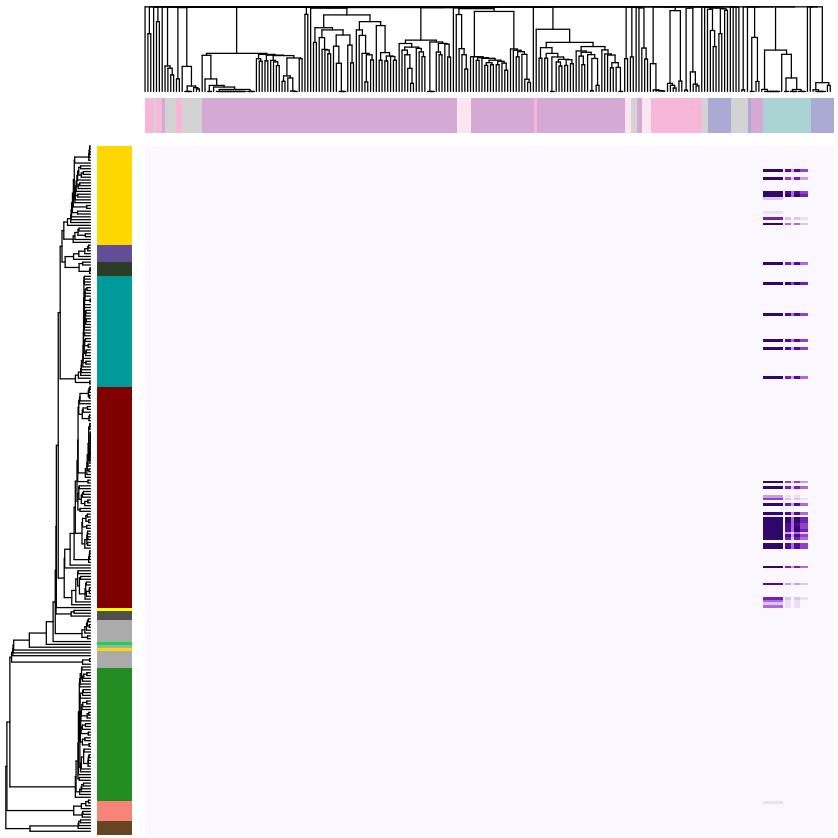


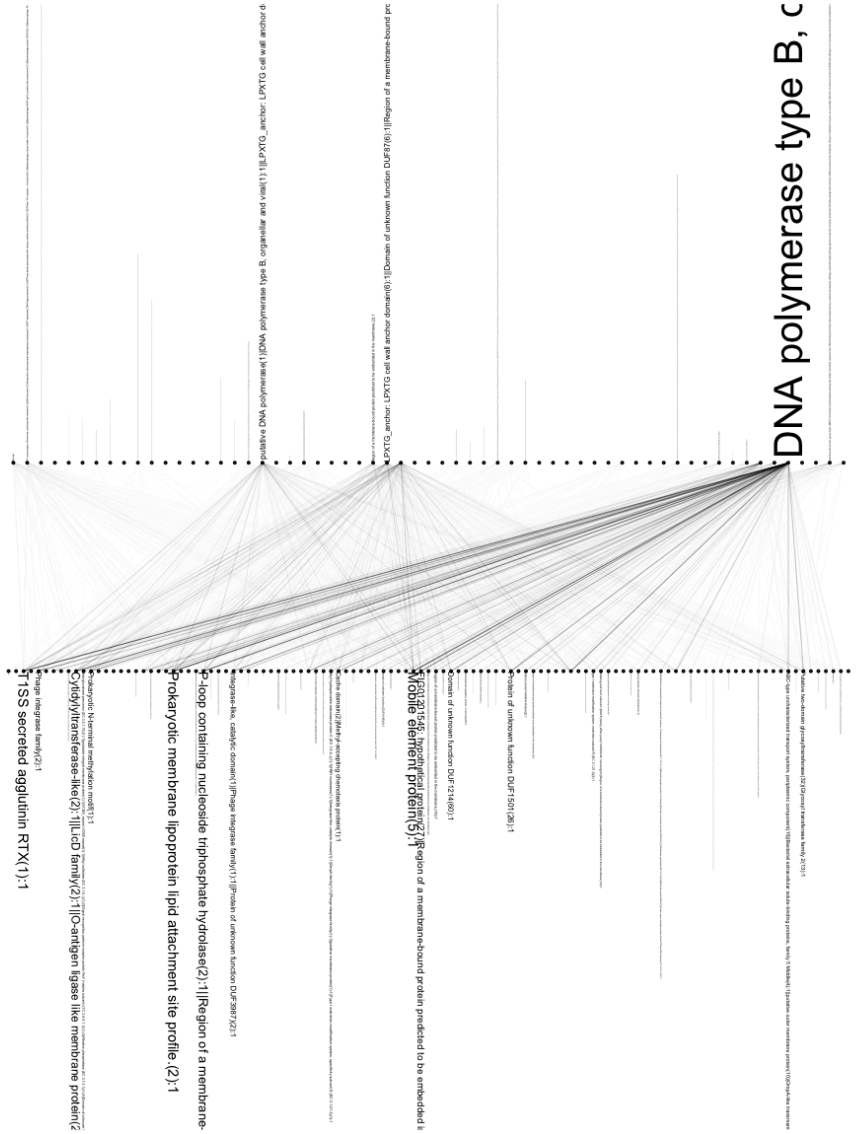
Orthogonal outer products may hint at interacting systems



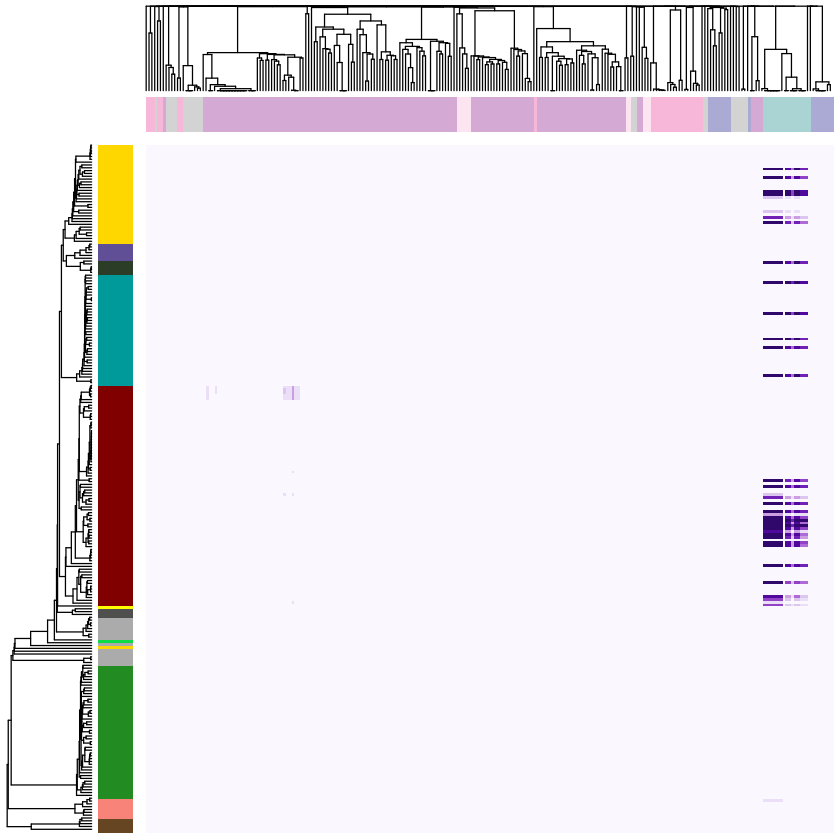


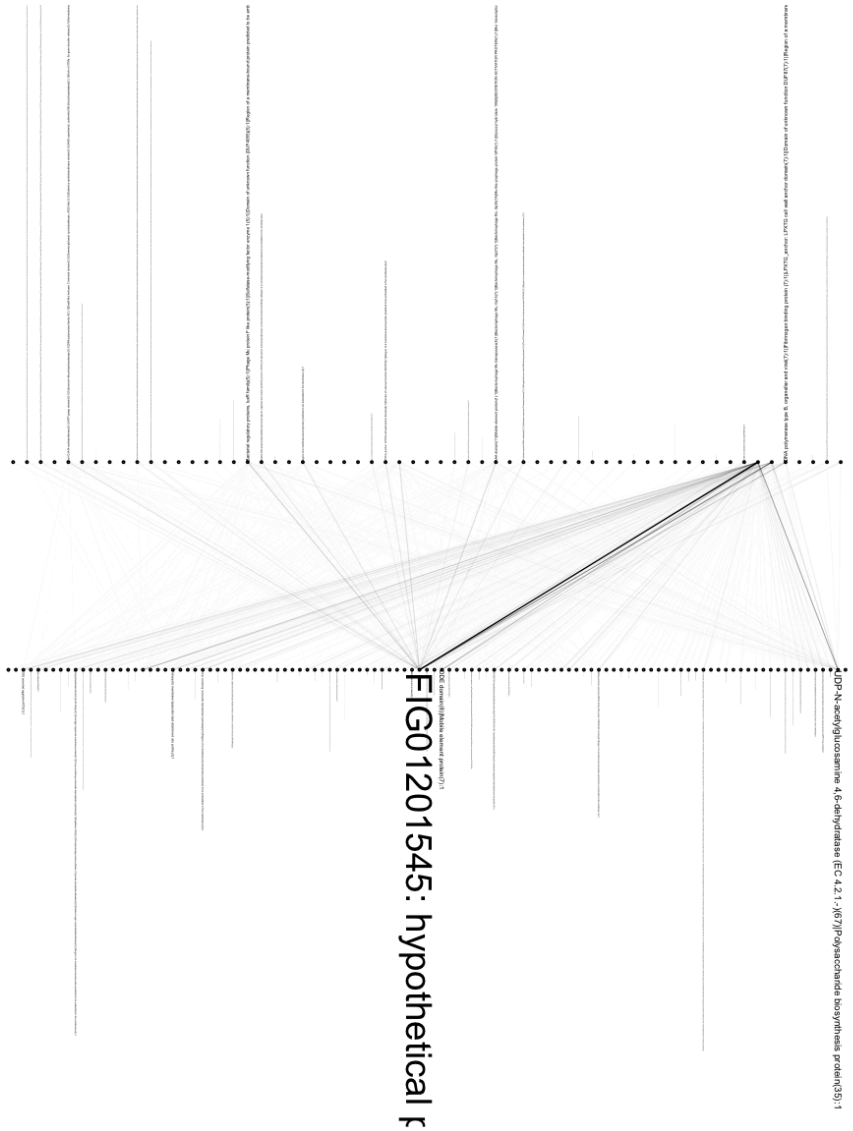
Orthogonal outer products may hint at interacting systems



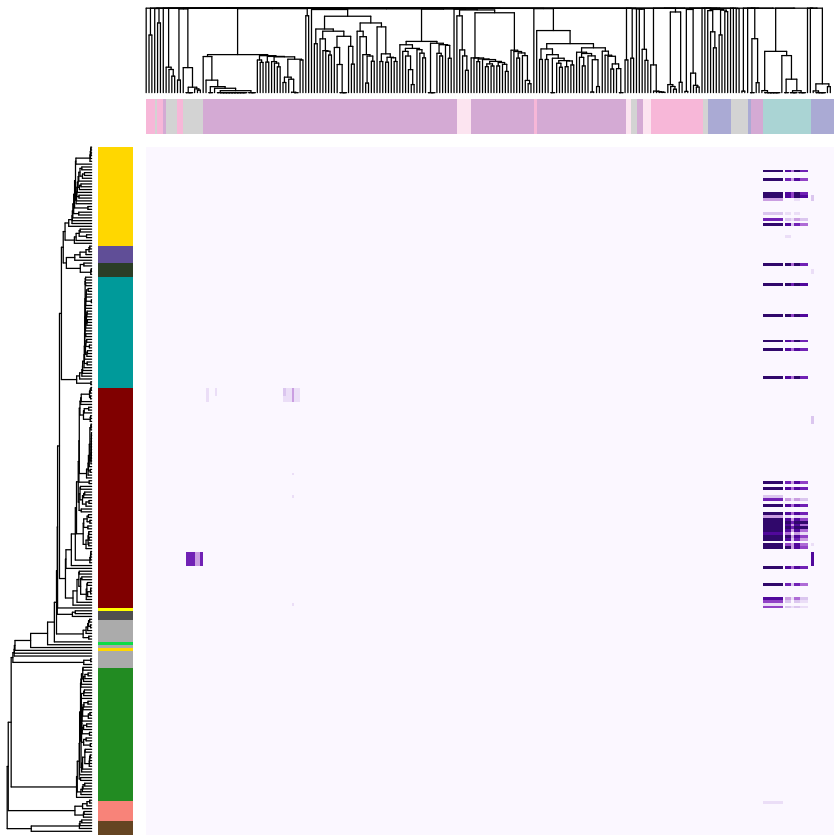


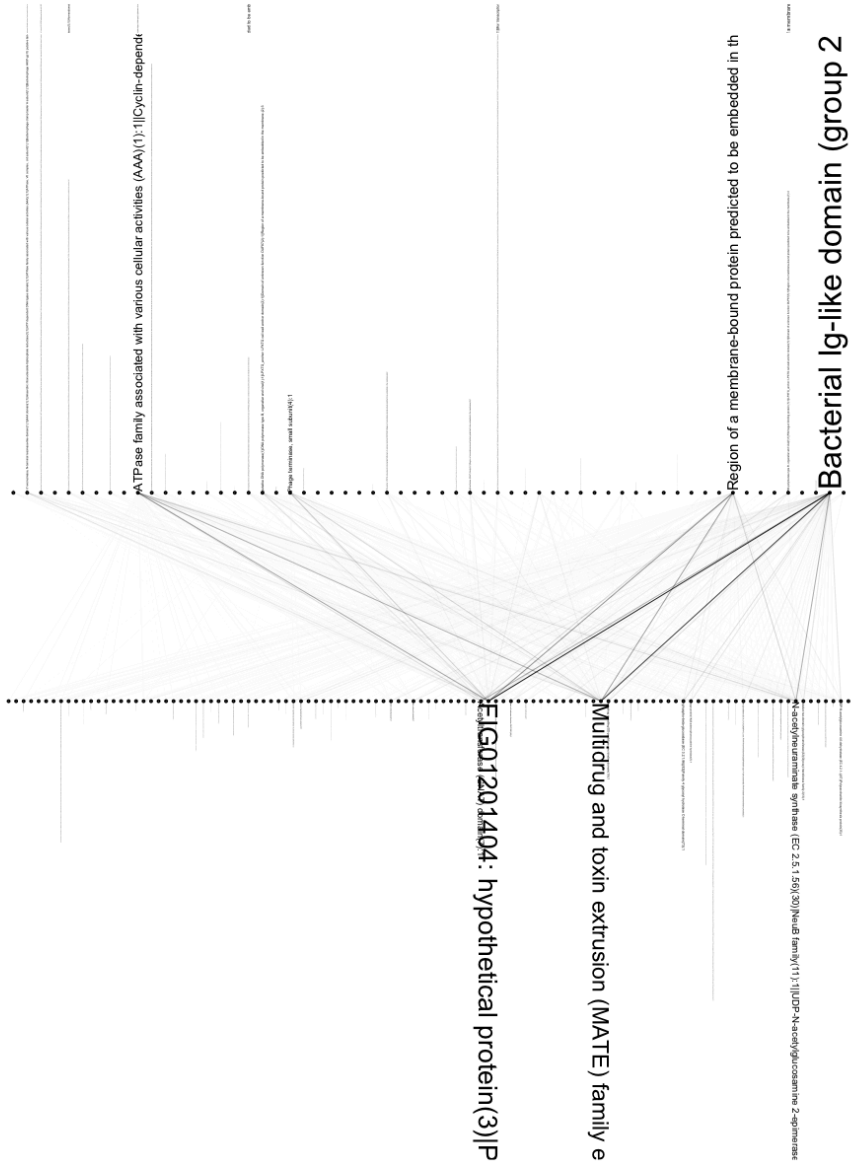
Orthogonal outer products may hint at interacting systems



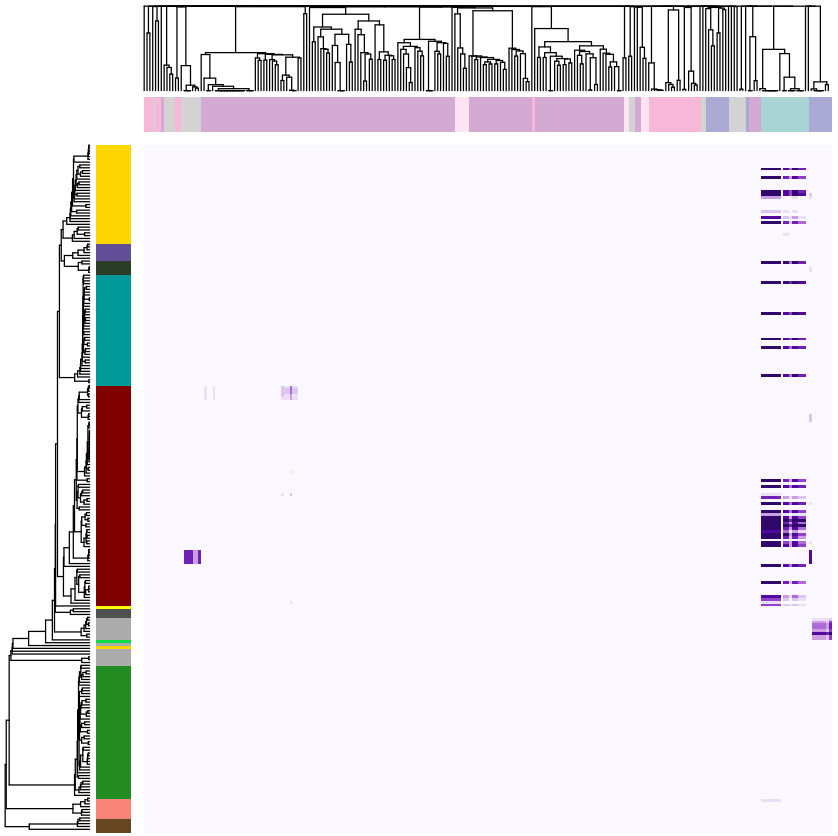


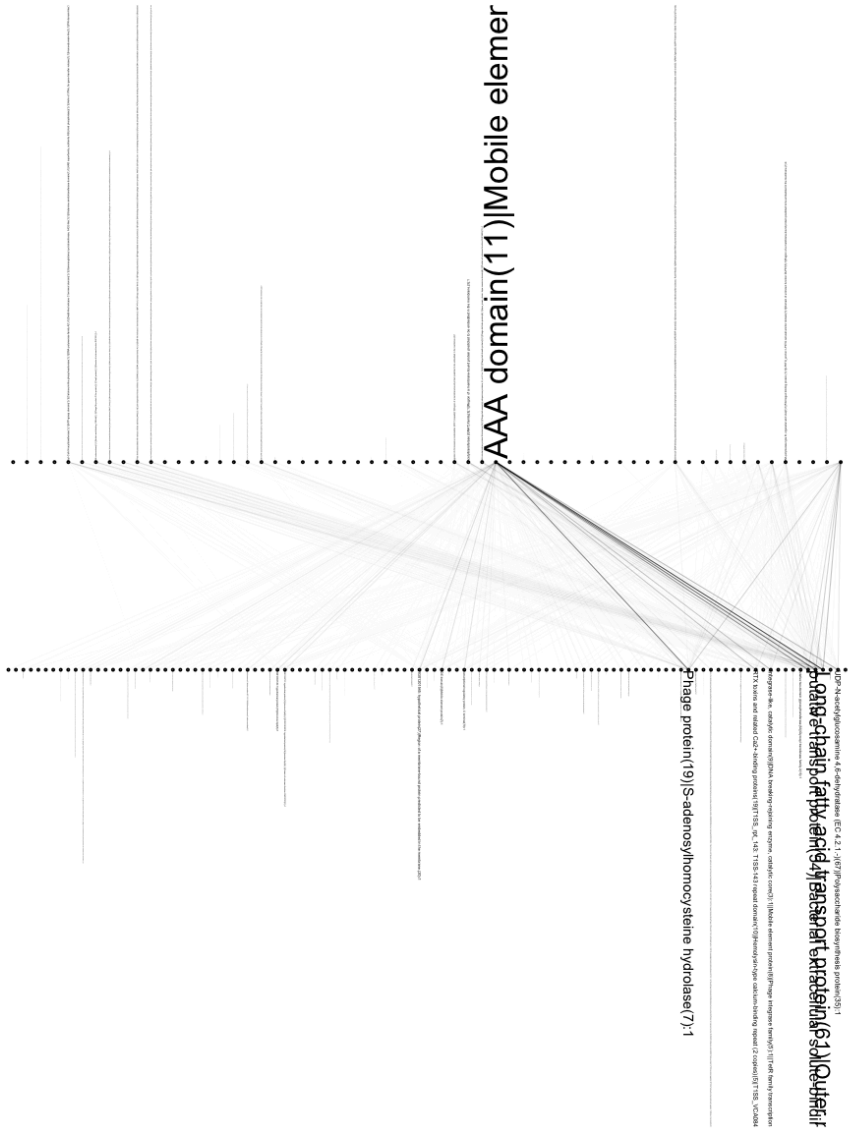
Orthogonal outer products may hint at interacting systems



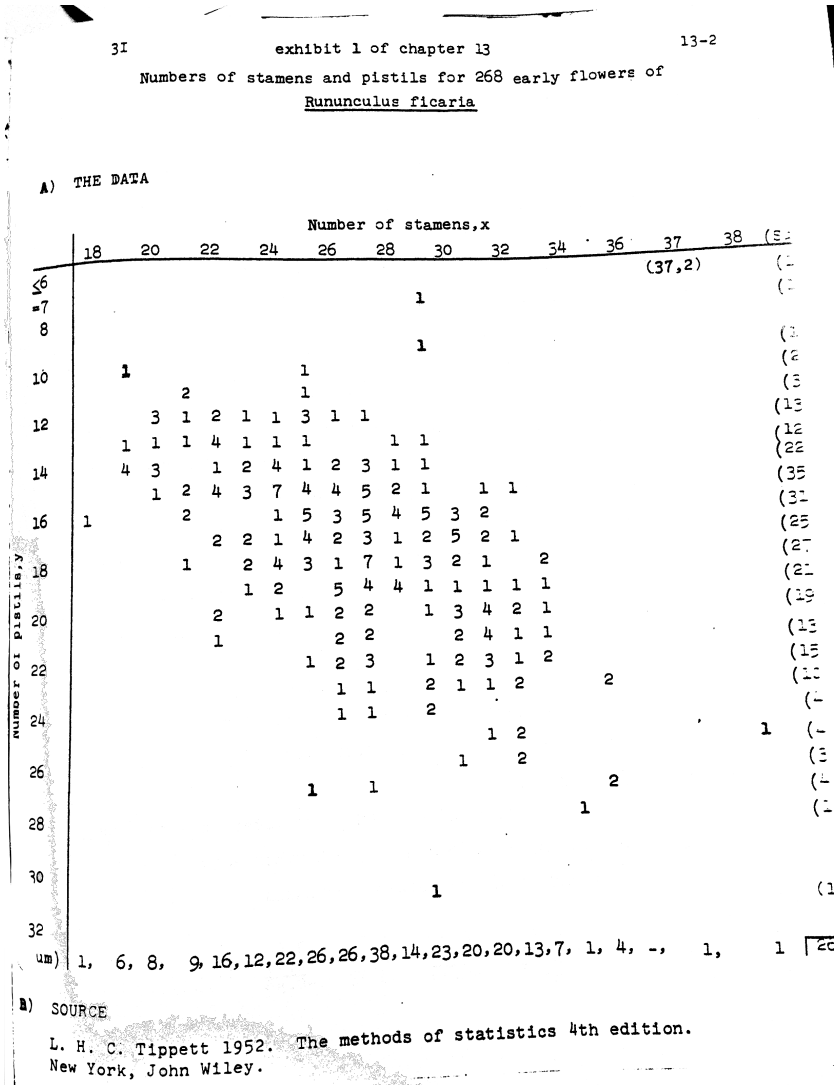


Orthogonal outer products may hint at interacting systems





Taking a step back, goal is to generate testable hypotheses, so we need still more intuitive ways to interact with the data



3I exhibit 1 of chapter 13 13-2

Numbers of stamens and pistils for 268 early flowers of Ranunculus ficaria

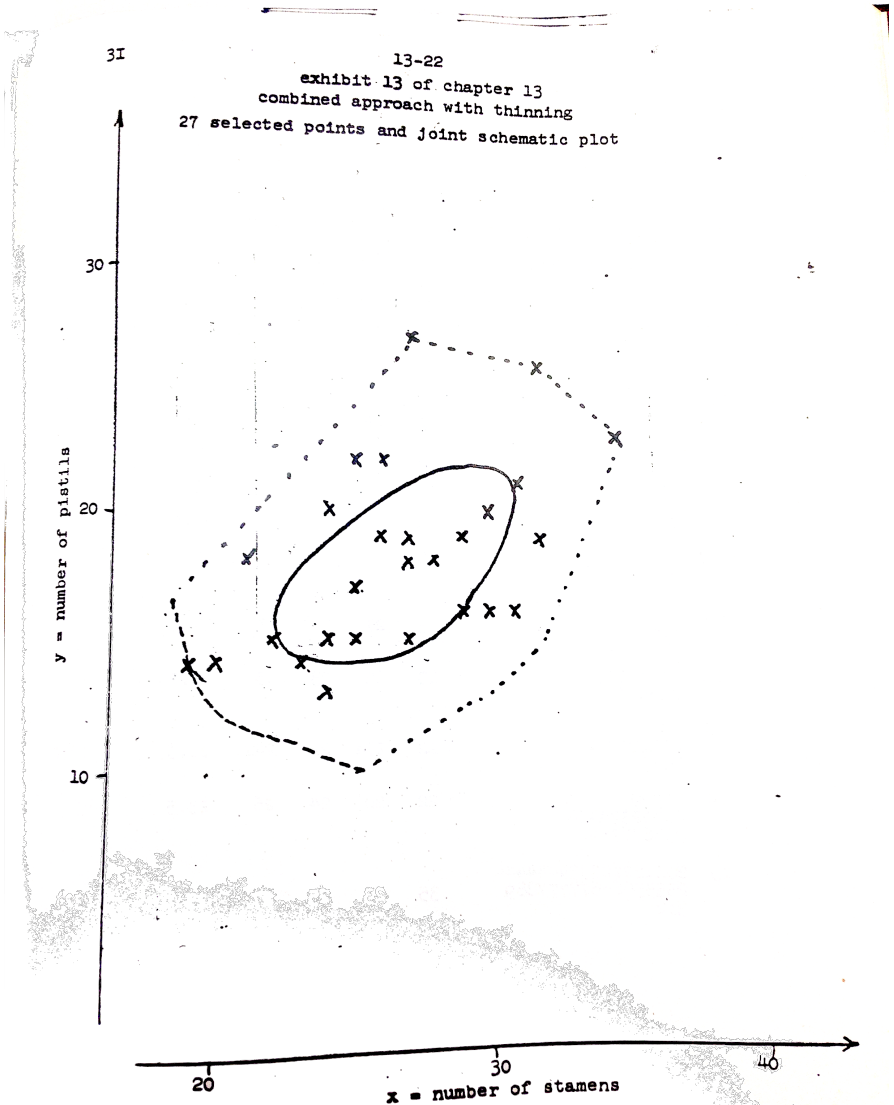
A) THE DATA

Number of stamens, x

number of pistils, y	18	20	22	24	26	28	30	32	34	36	37	38	(5
6											(37,2)		(-
=7							1						(-
8							1						(1
10		1			1								(2
			2		1								(3
12		3	1	2	1	1	3	1	1				(13
	1	1	1	4	1	1	1			1	1		(12
14	4	3		1	2	4	1	2	3	1	1		(22
		1	2	4	3	7	4	4	5	2	1		(35
16	1		2			1	5	3	5	4	5	3	(31
				2	2	1	4	2	3	1	2	5	(25
18			1		2	4	3	1	7	1	3	2	(27
					1	2		5	4	4	1	1	(21
20				2		1	1	2	2		1	3	(19
				1				2	2		2	4	(13
22							1	2	3		1	2	(15
								1	1		2	1	(11
24								1	1		2		(-
													1 (-
26											1		(3
					1		1						(-
28													(1
30													
							1						(1
32													
um)	1,	6,	8,	9,	16,	12,	22,	26,	26,	38,	14,	23,	20, 20, 13, 7, 1, 4, -, 1, 1 [26

B) SOURCE

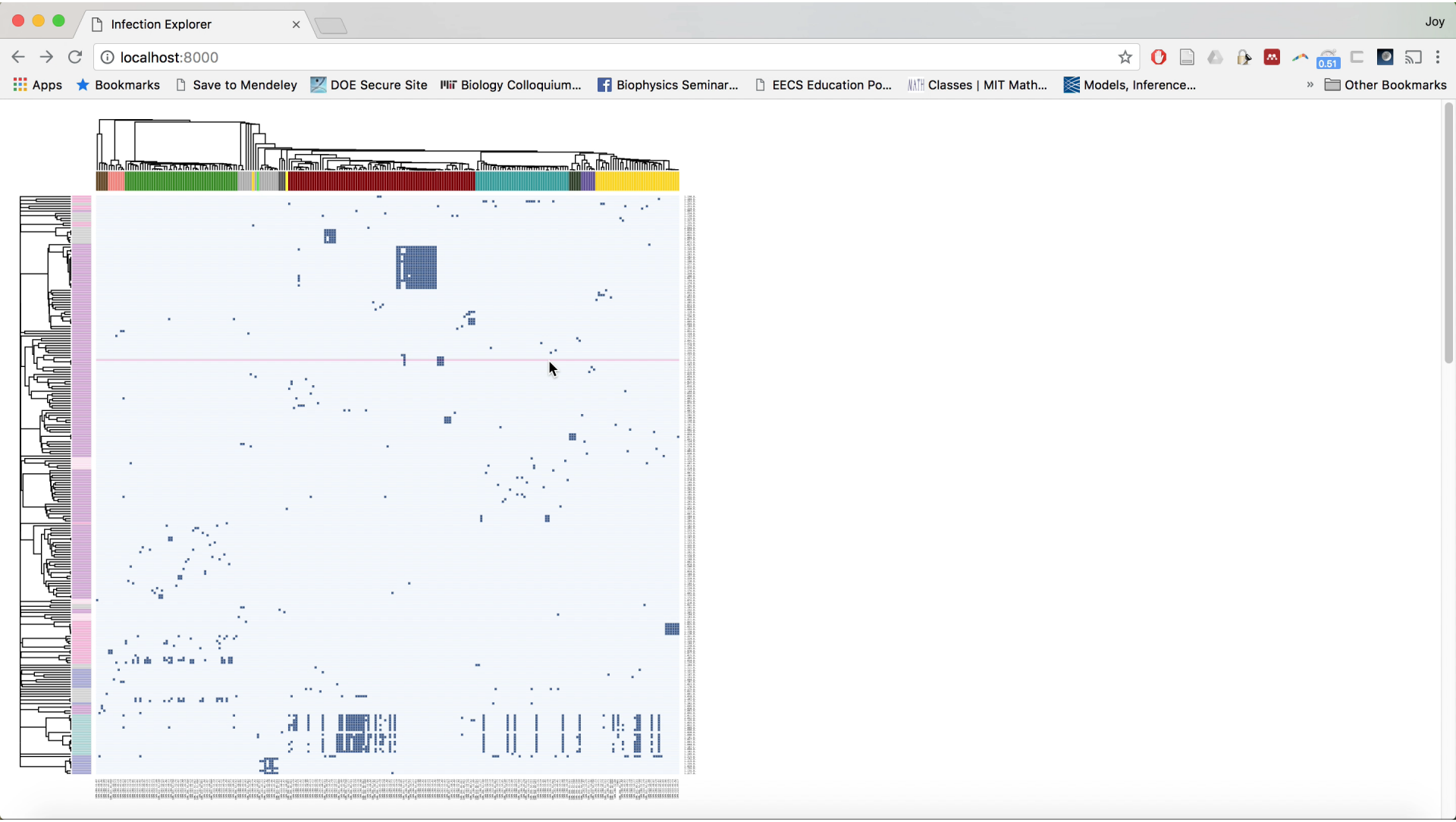
L. H. C. Tippett 1952. The methods of statistics 4th edition. New York, John Wiley.



3I 13-22

exhibit 13 of chapter 13
combined approach with thinning
27 selected points and joint schematic plot

Demo


Infection Explorer
Joy
localhost:8000
Apps
Bookmarks
Save to Mendeley
DOE Secure Site
Biology Colloquium...
Biophysics Seminar...
EECS Education Po...
Classes | MIT Math...
Models, Inference...
Other Bookmarks

Summary

- Correcting for phylogenetic confounding allows us to pick out defense mechanisms that would otherwise be lost among spurious correlations

Summary

- Correcting for phylogenetic confounding allows us to pick out defense mechanisms that would otherwise be lost among spurious correlations
- The multivariate model allows us to view the problem from a prediction perspective, and also helps us think about putative protein interactions between phage and host

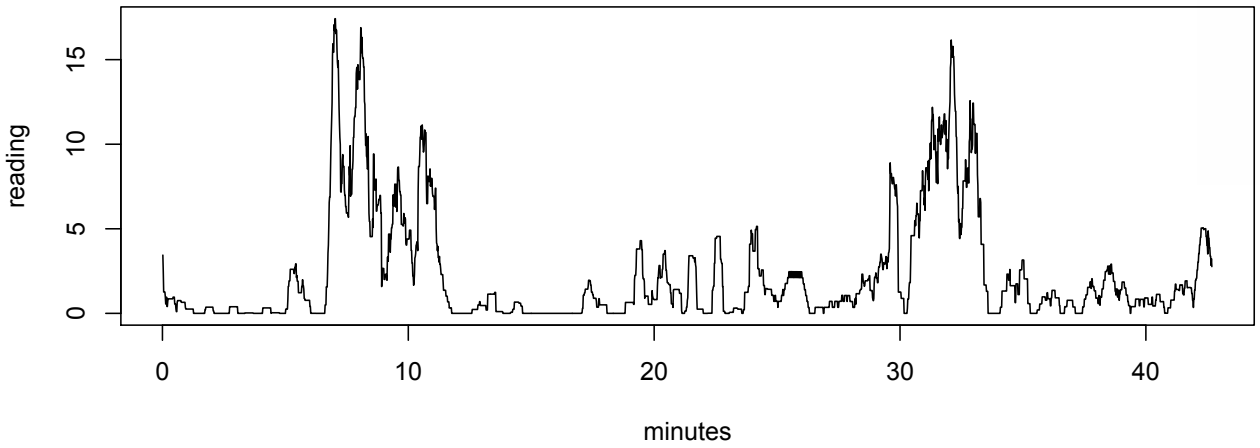
Summary

- Correcting for phylogenetic confounding allows us to pick out defense mechanisms that would otherwise be lost among spurious correlations
- The multivariate model allows us to view the problem from a prediction perspective, and also helps us think about putative protein interactions between phage and host
- Visualizing the data and results in an interactive manner, allows us to generate hypotheses about putative receptors and defense mechanisms

Aim 3: Developing modules for engineering/ statistical education

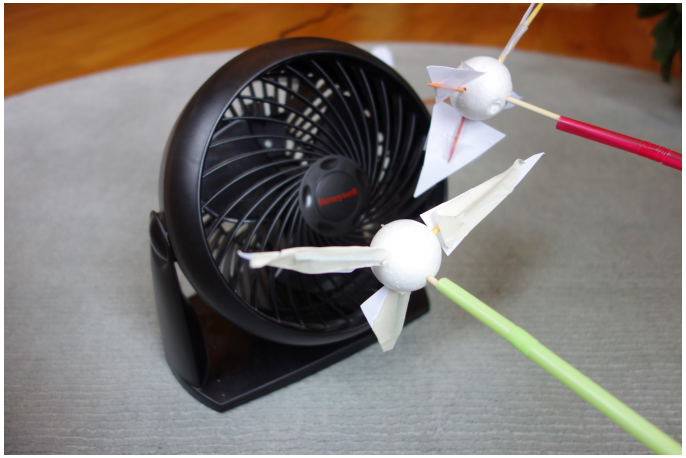
Incorporating statistics into Environmental Engineering for 8th Graders

- Building simple air quality sensors and interpreting the data collected (with Josh Moss/Kroll Lab)



Incorporating statistics into Environmental Engineering for 8^{th} Graders

- Building simple air quality sensors and interpreting the data collected (with Josh Moss/Kroll Lab)
- Experimental design applied to building wind turbines (with Ava Waitz)



Incorporating statistics into Environmental Engineering for 8th Graders

- Building simple air quality sensors and interpreting the data collected (with Josh Moss/Kroll Lab)
- Experimental design applied to building wind turbines (with Ava Waitz)
- Interpreting maps – solar irradiance, rainfall, soil types, crop yields
- Population modeling using game theory – prisoner's dilemma/rock-paper-scissors/etc
- Etc. (lots of guidance from Anjuli Jain)

Thank You

Advisors

Martin Polz
Libusha Kelly

Polz Lab

Kathryn Kauffman
Fatima Hussain
David VanInsberghe

Joseph Elsherbini
Annie Yu
Fabiola Miranda

Javier Dubert
Bruno Janeiro
Clovis Borges

MIT Parsons

Anjuli Jain
+ Really, all of Parsons

Committee

David Bartel
Philippe Rigollet
Jeff Gore

Practicum Lab

Adam Arkin
Harneet Rishi