Which Phenotypes Affect Bacteria's Inhibition Ability?

Yang Yang, Cheng Zhang

University of Minnesota

September 12, 2014

Who and When

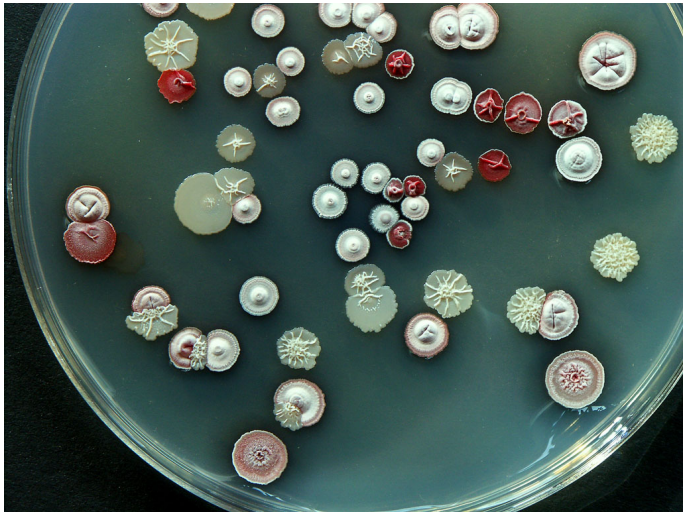
- Summer of 2014, consulting project through School of Statistics consulting clinic

Who and When

- Summer of 2014, consulting project through School of Statistics consulting clinic
- Linda Kinkel: Department of Plant Pathology
- Cheng Zhang
- Yang Yang
- Aaron Renhdal

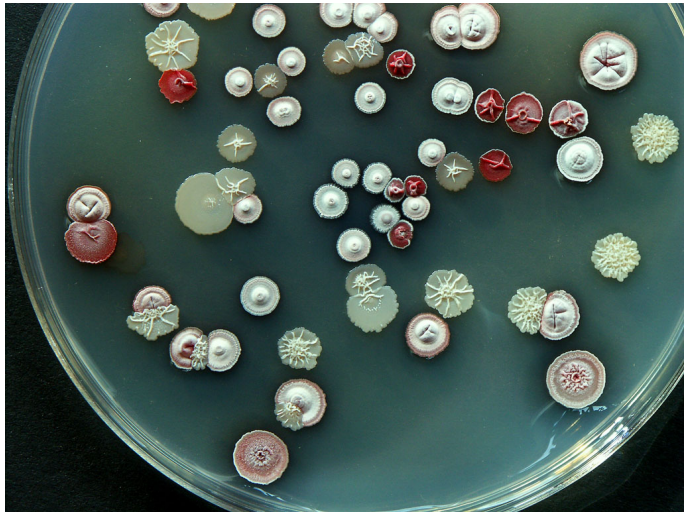
What

- Bacteria can produce antibiotics



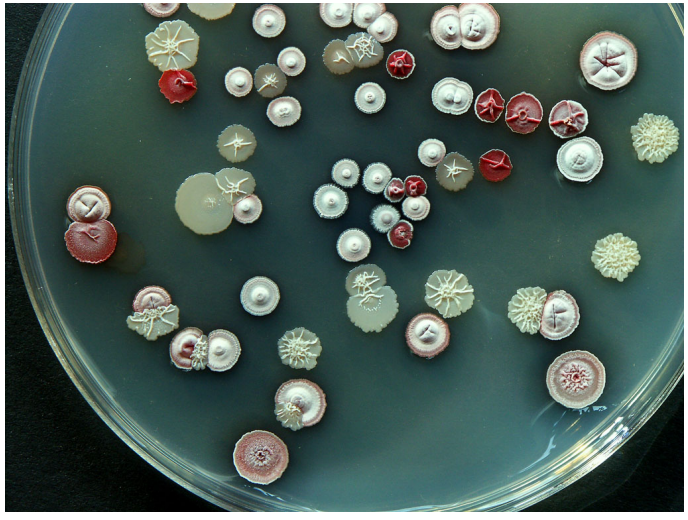
What

- Bacteria can produce antibiotics
- Our research subject: *streptomyces*



What

- Bacteria can produce antibiotics
- Our research subject: *streptomyces*
- How does location, genetic similarity and niche overlap affect *streptomyces*'s inhibitory ability?



Why

- *Streptomyces* is the largest antibiotic-producing genus

Why

- *Streptomyces* is the largest antibiotic-producing genus
- Streptomycin was the first cure for tuberculosis

Why

- *Streptomyces* is the largest antibiotic-producing genus
- Streptomycin was the first cure for tuberculosis
- The mechanism of antibiotics production remains unanswered

The Data

- Global data: 83 different *Streptomyces* isolates, 7 locations around the world

The Data

- Global data: 83 different *Streptomyces* isolates, 7 locations around the world
- We were not involved in data collection nor designing experiment

Response Variable: Size of the Killing Zone

- Grow isolate A on agar plate, A will produce antibiotics

Response Variable: Size of the Killing Zone

- Grow isolate A on agar plate, A will produce antibiotics
- Kill all isolate A, then grow isolate B on the same plate

Response Variable: Size of the Killing Zone

- Grow isolate A on agar plate, A will produce antibiotics
- Kill all isolate A, then grow isolate B on the same plate
- Measure the size of the area where B does not grow. Then this area is the killing ability of $A \rightarrow B$

Response Variable: Size of the Killing Zone

- Grow isolate A on agar plate, A will produce antibiotics
- Kill all isolate A, then grow isolate B on the same plate
- Measure the size of the area where B does not grow. Then this area is the killing ability of $A \rightarrow B$
- Note: $A \rightarrow B \neq B \rightarrow A$

Response Variable: Size of the Killing Zone

- Grow isolate A on agar plate, A will produce antibiotics
- Kill all isolate A, then grow isolate B on the same plate
- Measure the size of the area where B does not grow. Then this area is the killing ability of $A \rightarrow B$
- Note: $A \rightarrow B \neq B \rightarrow A$
- Stored as an asymmetric matrix

Response Variable: Size of the Killing Zone

- Grow isolate A on agar plate, A will produce antibiotics
- Kill all isolate A, then grow isolate B on the same plate
- Measure the size of the area where B does not grow. Then this area is the killing ability of $A \rightarrow B$
- Note: $A \rightarrow B \neq B \rightarrow A$
- Stored as an asymmetric matrix
- Need to do $P_2^{83} = 6806$ experiments!

Predictor Variable: Niche Overlap

- Measures the proportion of common nutrients between two isolates

Predictor Variable: Niche Overlap

- Measures the proportion of common nutrients between two isolates
- Grow each isolate in 95 different nutrients

Predictor Variable: Niche Overlap

- Measures the proportion of common nutrients between two isolates
- Grow each isolate in 95 different nutrients
- Example: A uses 40 units of nutrients in total, B uses 20 units of nutrients in total, 10 units of nutrients are overlapping. Then niche overlap for A is: 25%, for B is 50%

Predictor Variable: Niche Overlap

- Measures the proportion of common nutrients between two isolates
- Grow each isolate in 95 different nutrients
- Example: A uses 40 units of nutrients in total, B uses 20 units of nutrients in total, 10 units of nutrients are overlapping. Then niche overlap for A is: 25%, for B is 50%
- We wrote `R` function to calculate the niche overlap from raw data and store the values in a square matrix (NOT symmetric)

Predictor Variable: Genetic Distance

- Between 0-1
- Measures the similarity between the genes of two isolates
- Calculated by Biology Workbench
- Stored as a symmetric matrix

Predictor Variable: Locations

- Locations were treated as factors

Preliminary Work

- Clients claim that they have found significant correlation between "certain" predictors and response variable

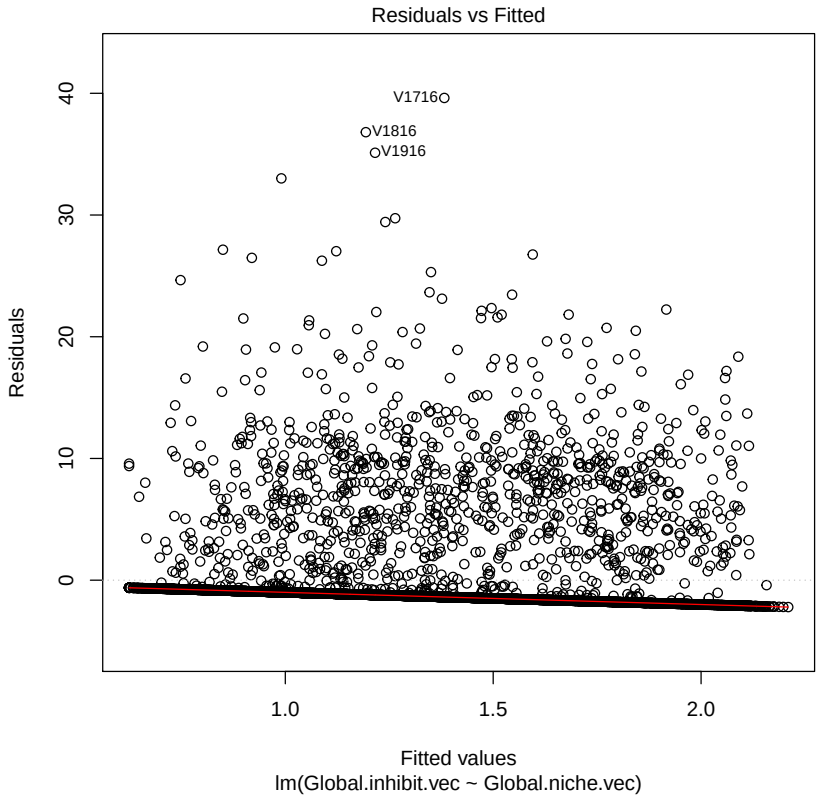
Preliminary Work

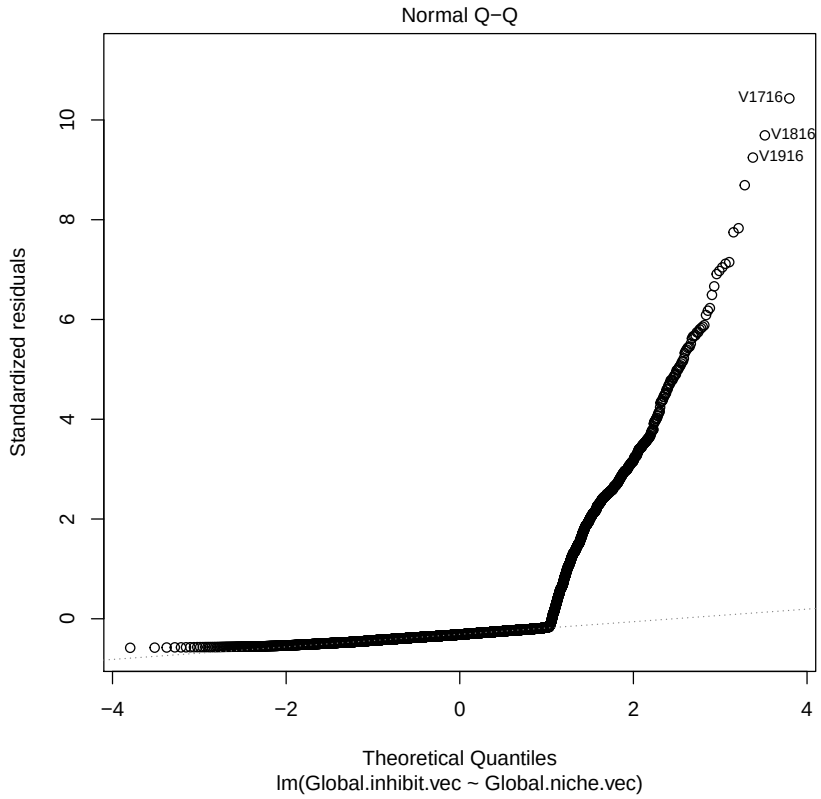
- Clients claim that they have found significant correlation between "certain" predictors and response variable
- We started with multiple linear regression, tried to reproduce clients' results

However...

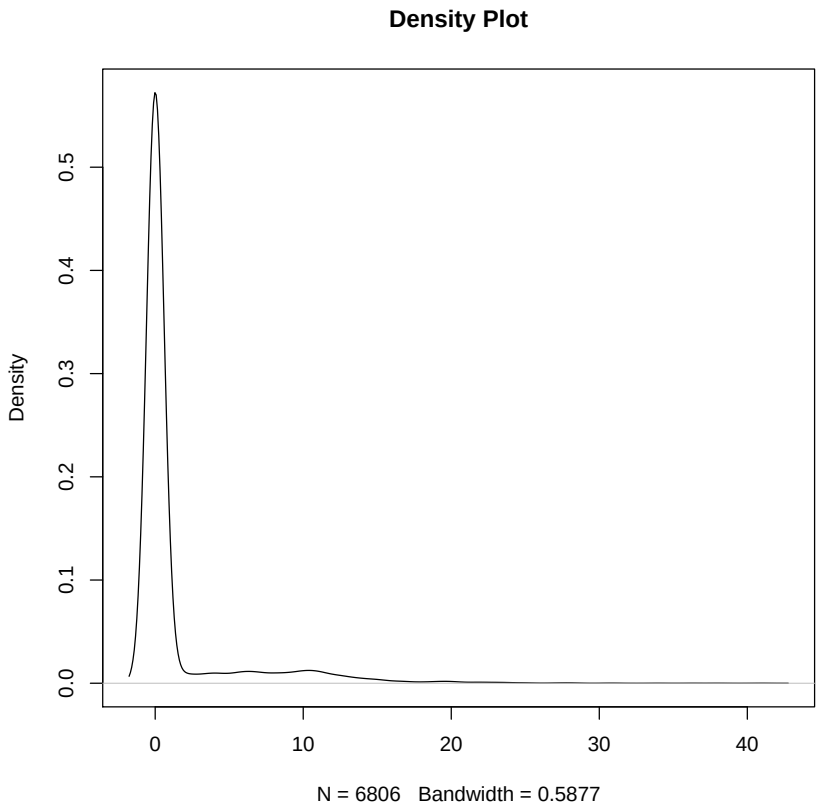
- We could not reproduce the same results
- Multiple regression did not fit the data well

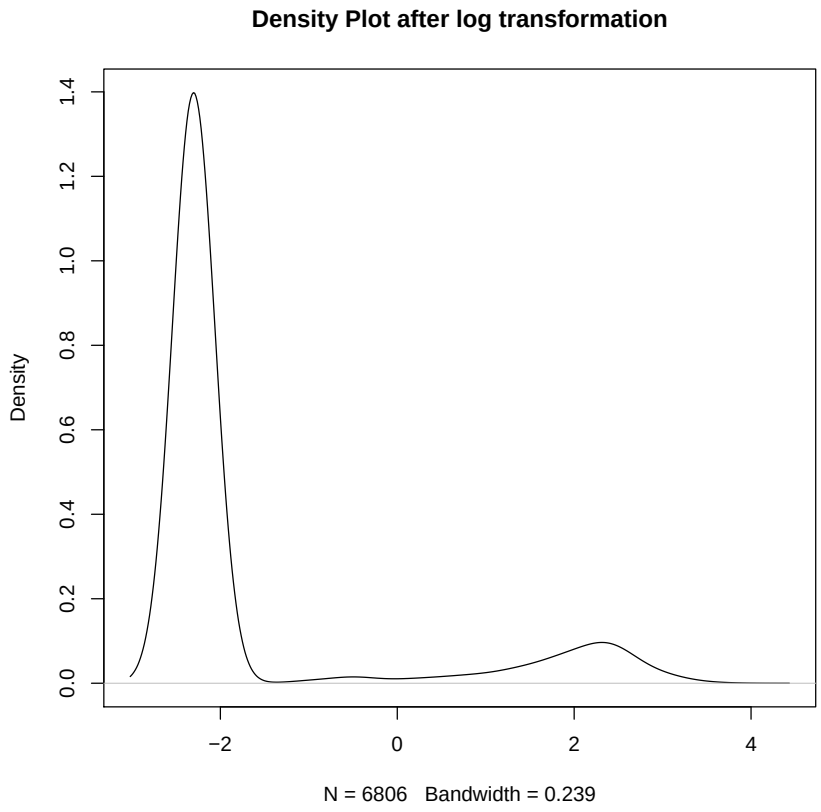
Regression Diagnostics





Zero Inflated Response





83% of the response are zeros!

Inspiration: Auto Insurance Data

- Similar to modeling auto insurance premium

Inspiration: Auto Insurance Data

- Similar to modeling auto insurance premium
- In auto insurance data, over 90% of the data does not have any claim. (response$= 0$)

Inspiration: Auto Insurance Data

- Similar to modeling auto insurance premium
- In auto insurance data, over 90% of the data does not have any claim. (response$= 0$)
- Zero Adjusted Inverse Gaussian (ZAIG) model has been well established

ZAIG in a nutshell

- Model the probability of killing and killing ability separately

ZAIG in a nutshell

- Model the probability of killing and killing ability separately
- Killing$\sim$Bernoulli$(1, \pi)$

ZAIG in a nutshell

- Model the probability of killing and killing ability separately
- Killing$\sim$Bernoulli$(1, \pi)$
- Size of killing zone$\sim$Inverse Gaussian(μ, σ)

ZAIG in a nutshell

- Model the probability of killing and killing ability separately
- Killing$\sim$Bernoulli$(1, \pi)$
- Size of killing zone$\sim$Inverse Gaussian(μ, σ)
- Mixed continuous-discrete distribution

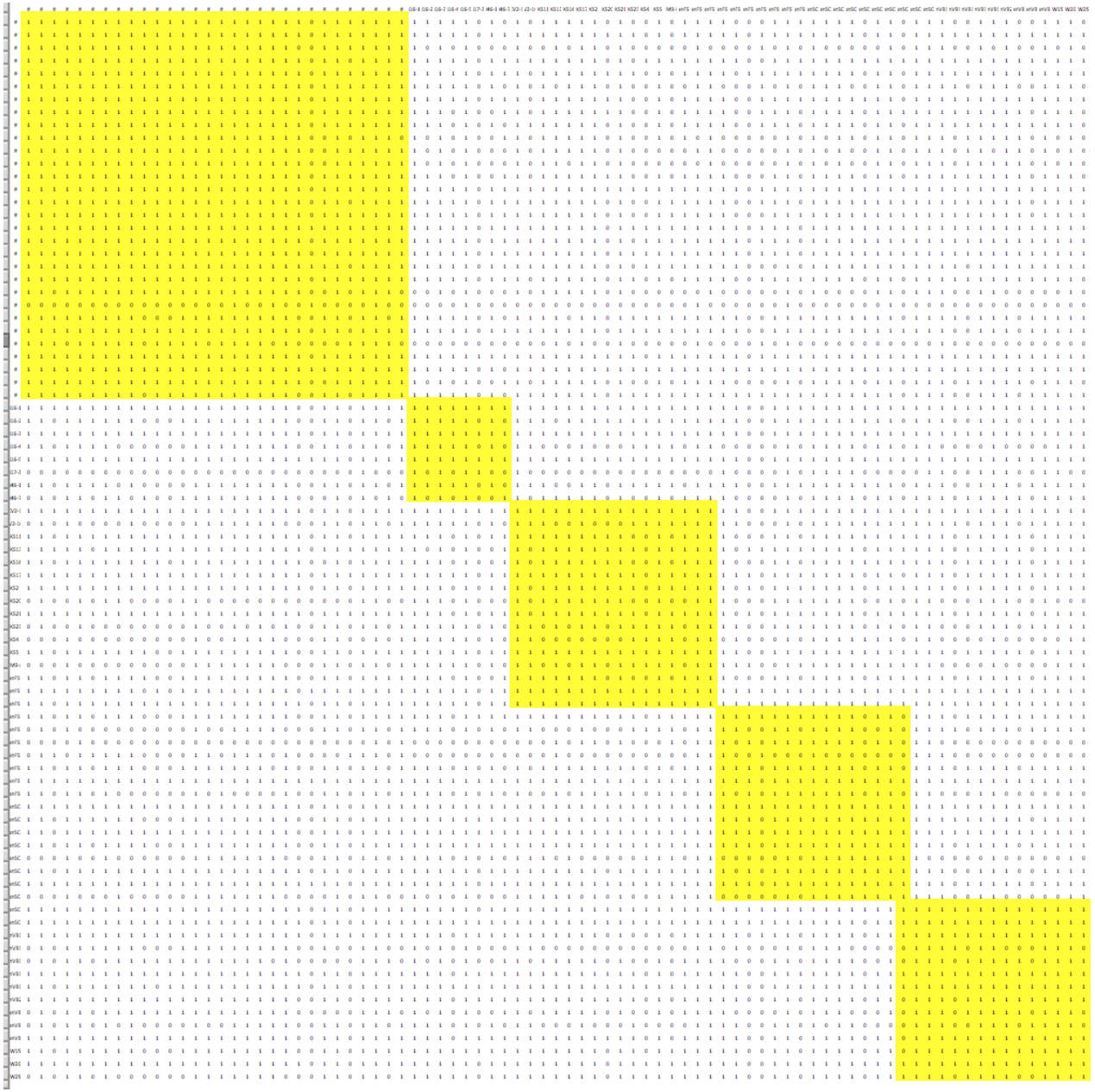
ZAIG in a nutshell

- Model the probability of killing and killing ability separately
- Killing$\sim$Bernoulli$(1, \pi)$
- Size of killing zone$\sim$Inverse Gaussian(μ, σ)
- Mixed continuous-discrete distribution
- `library(gamlss)`

Sympatric Analysis

- Clients specifically asked for two models
- Sympatric: intra-location
- Allopatric: inter-location

Sympatric Analysis



Sympatric Analysis

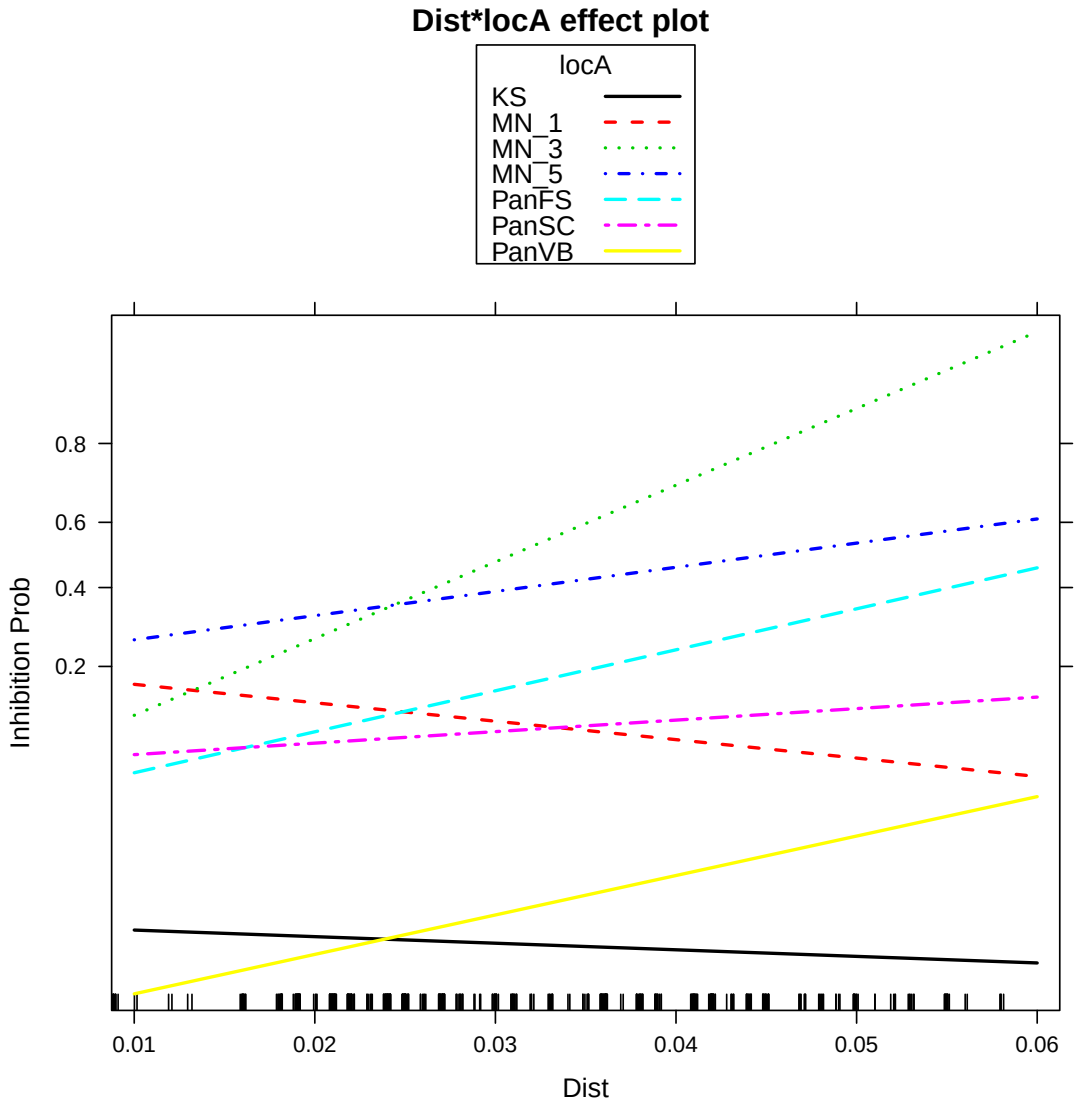


Figure : Interaction of `Dist:locA`

Sympatric Analysis

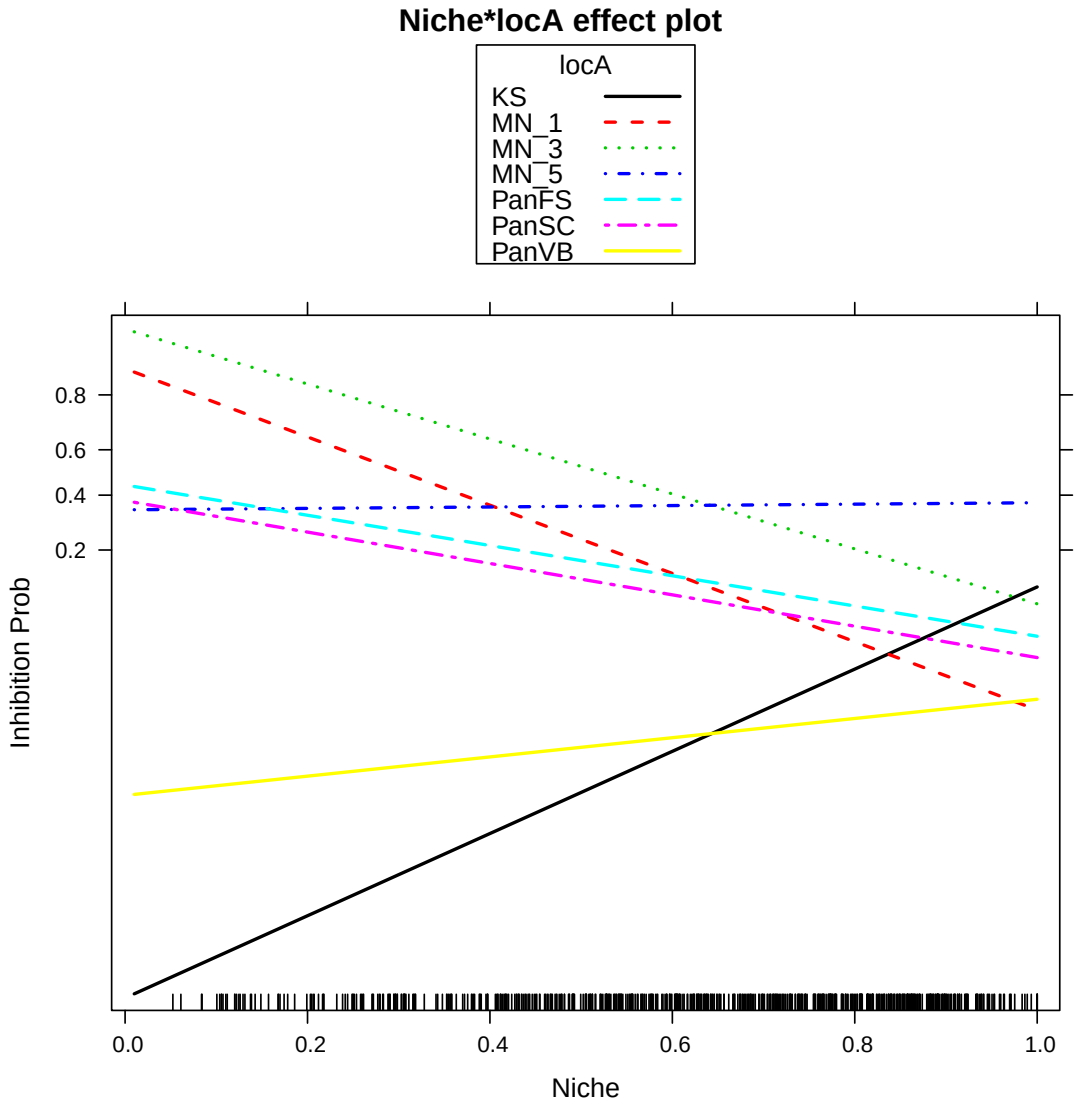
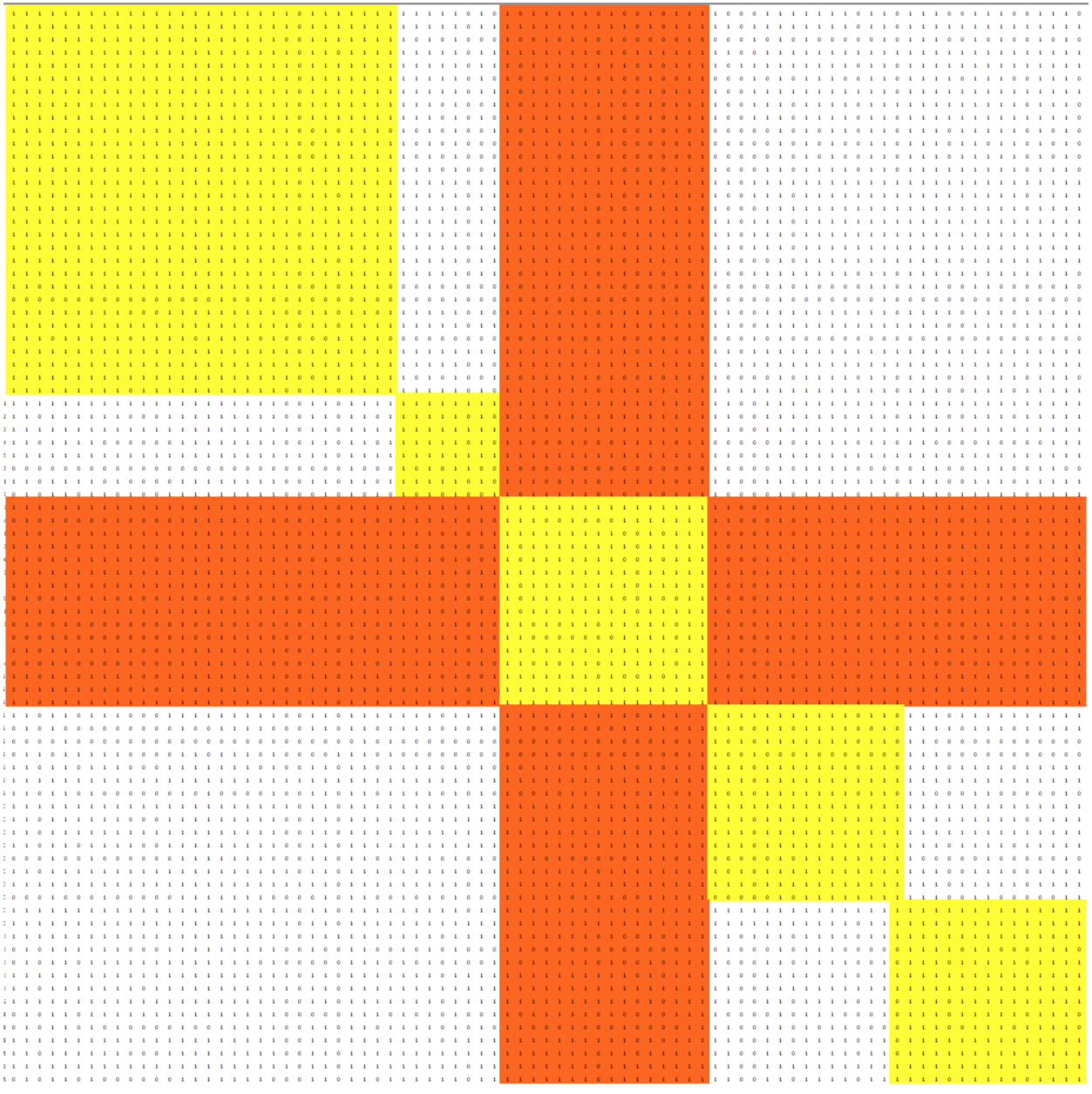


Figure : Interaction of `Niche:locA`

Allopatric Analysis



Allopatric Analysis: MN1

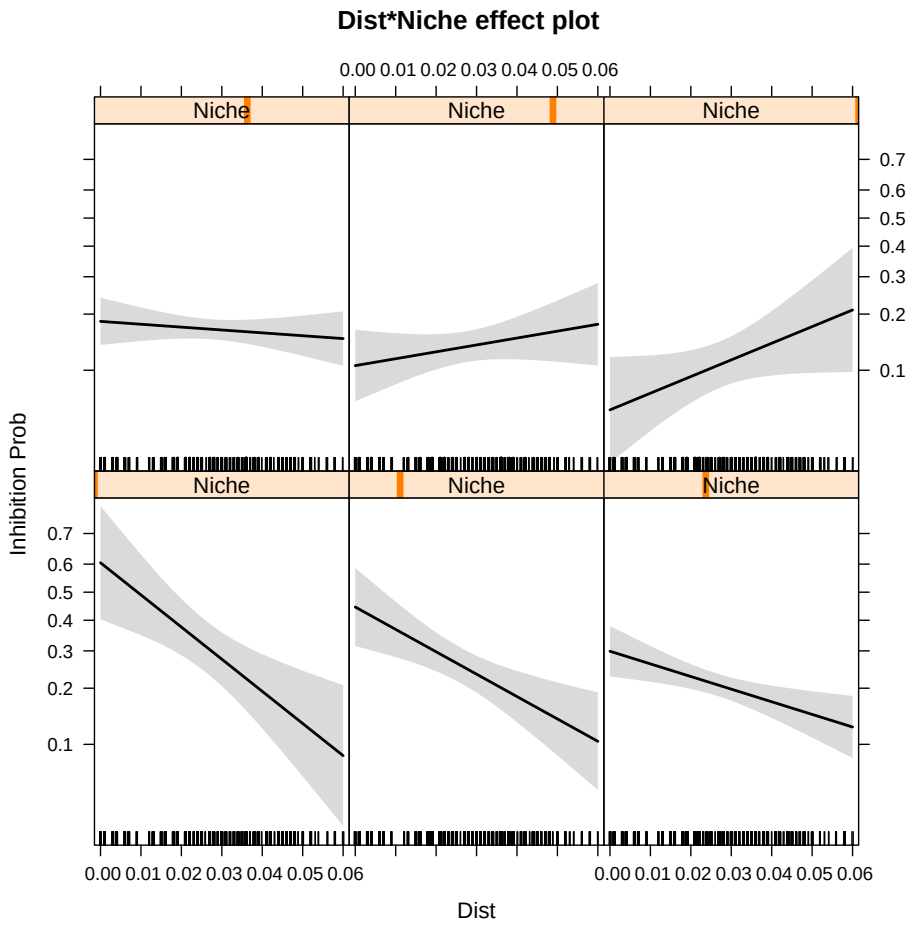


Figure : Interaction plot of `Dist:Niche` at MN1

- Inhibition size: none of the predictors are significant.

Allopatric Analysis: MN3

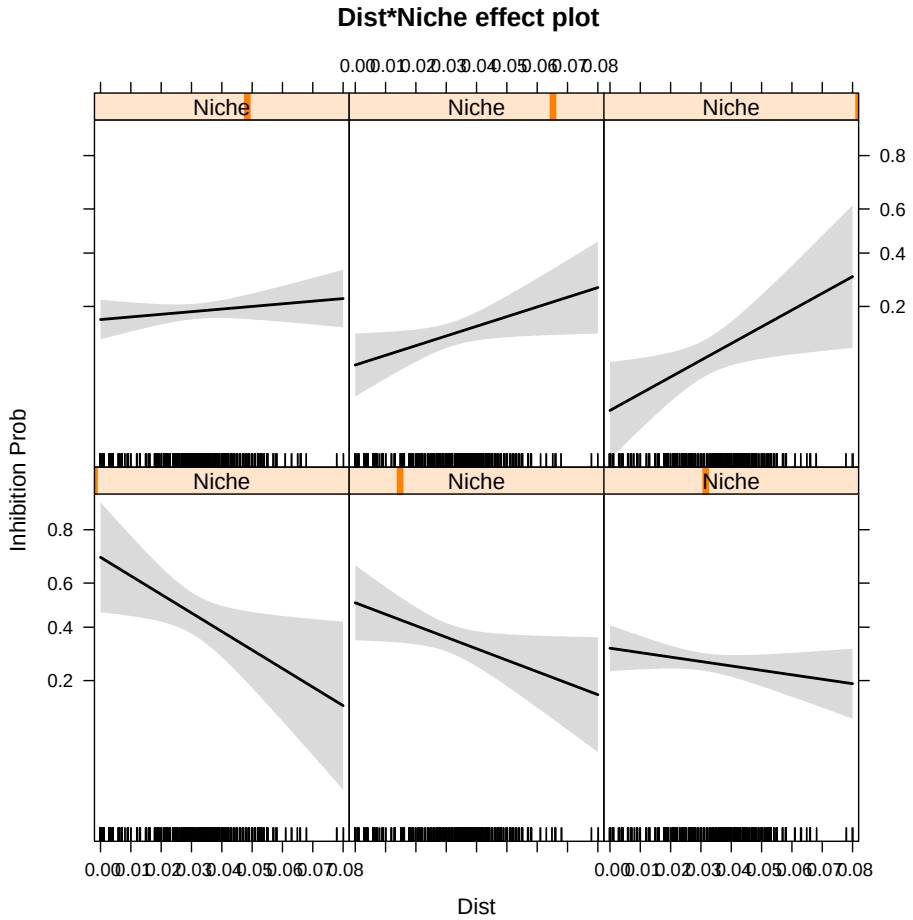


Figure : Interaction plot of `Dist:Niche` at MN3

- Inhibition size: genetic distance is significant ($p = 0.029$). The inhibition size increases 8.8% as the distance increases by 0.01 unit.

Allopatric Analysis: MN5

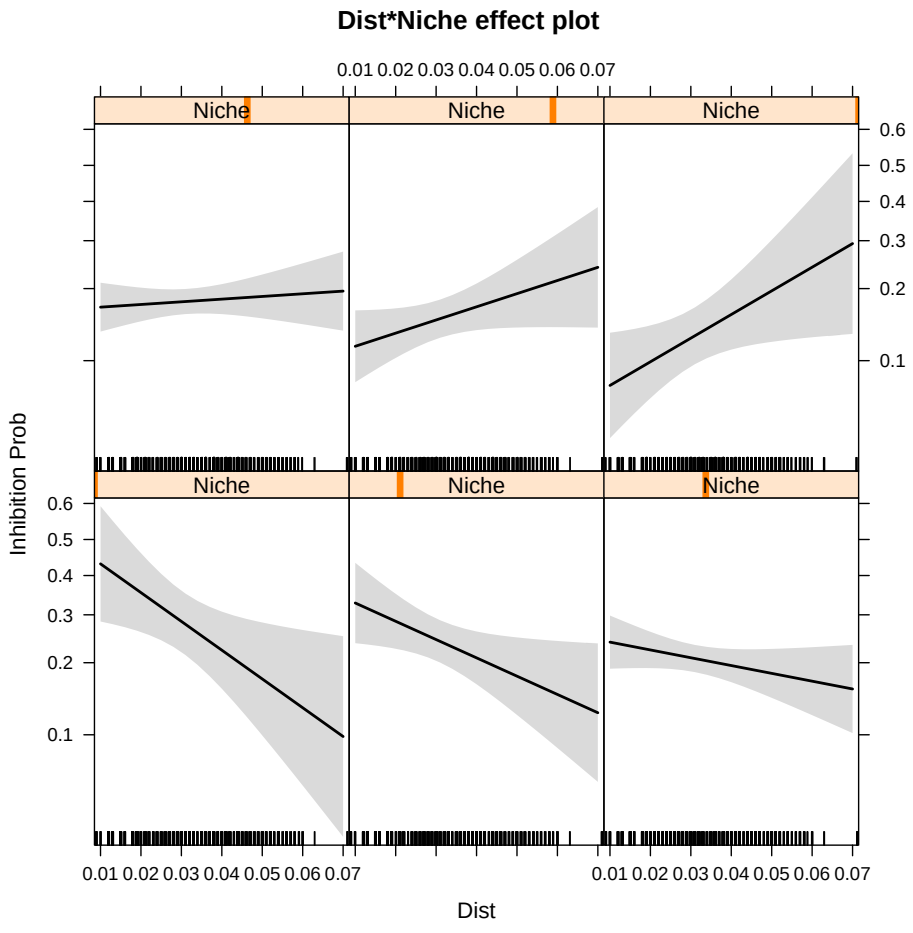


Figure : Interaction plot of `Dist:Niche` at MN5

- Inhibition size: none of the predictors are significant

Allopatric Analysis: Kansas

- Inhibition probability: none of the predictors are significant.
- Inhibition size: none of the predictors are significant.

Allopatric Analysis: PanFS

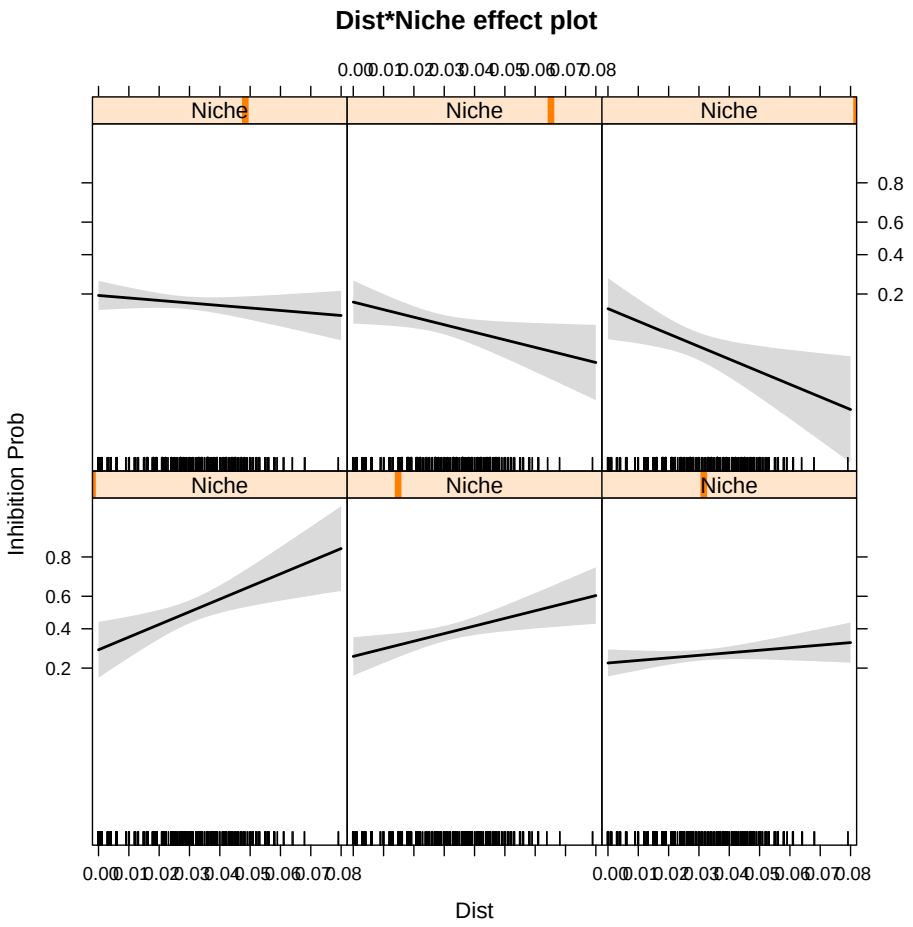


Figure : Interaction plot of `Dist:Niche` at PanFS

- Inhibition size: none of the predictors are significant

Allopatric Analysis: PanSC

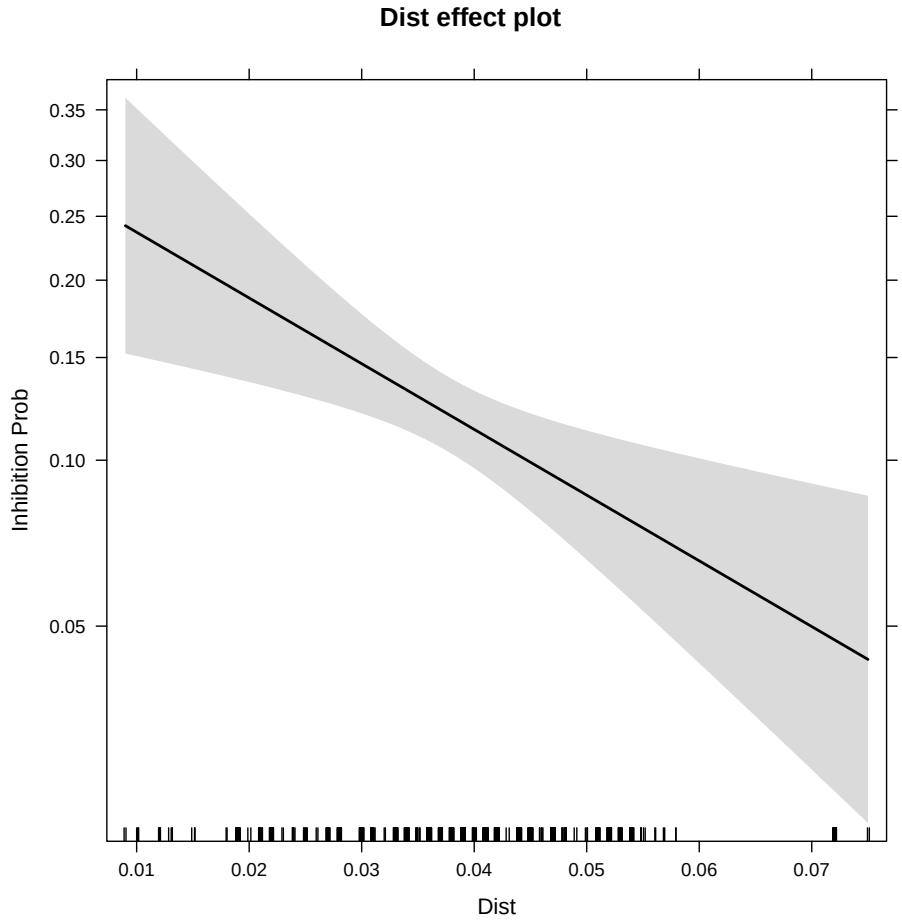


Figure : Effect plot of `Dist` at PanSC

- Inhibition size: Niche overlap is significant ($p \approx 0$). As `Niche` increases by 0.1 unit, inhibition size decreases 11.7%.

Allopatric Analysis: PanVB

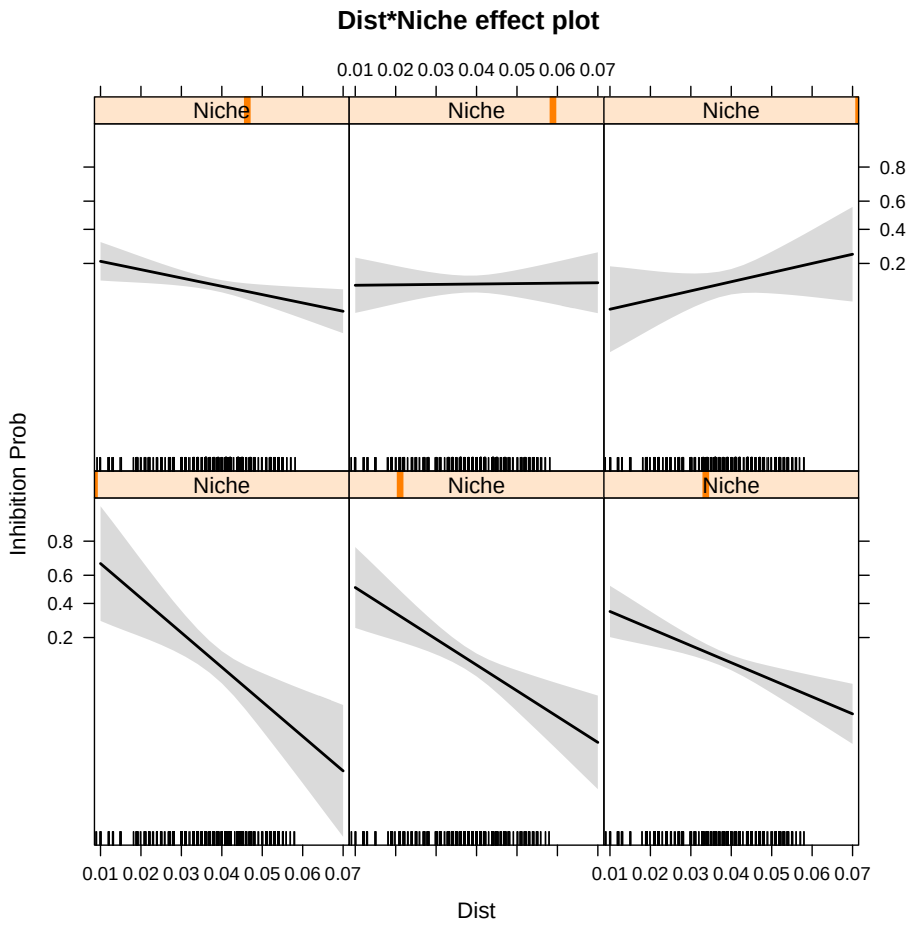


Figure : Interaction plot of `Dist:Niche` at PanVB

- Inhibition size: none of the predictors are significant.

Take home message

- How to interpret the results?

Take home message

- How to interpret the results?
- Be careful with the “prior” information provided by clients

Take home message

- How to interpret the results?
- Be careful with the “prior” information provided by clients
- e.g., pre-processed data, preliminary analysis

Thank you

Questions?