UNIVERSITY OF SAN FRANCISCO

Nonabelian Fourier Transforms: A Path to Solving Epistasis

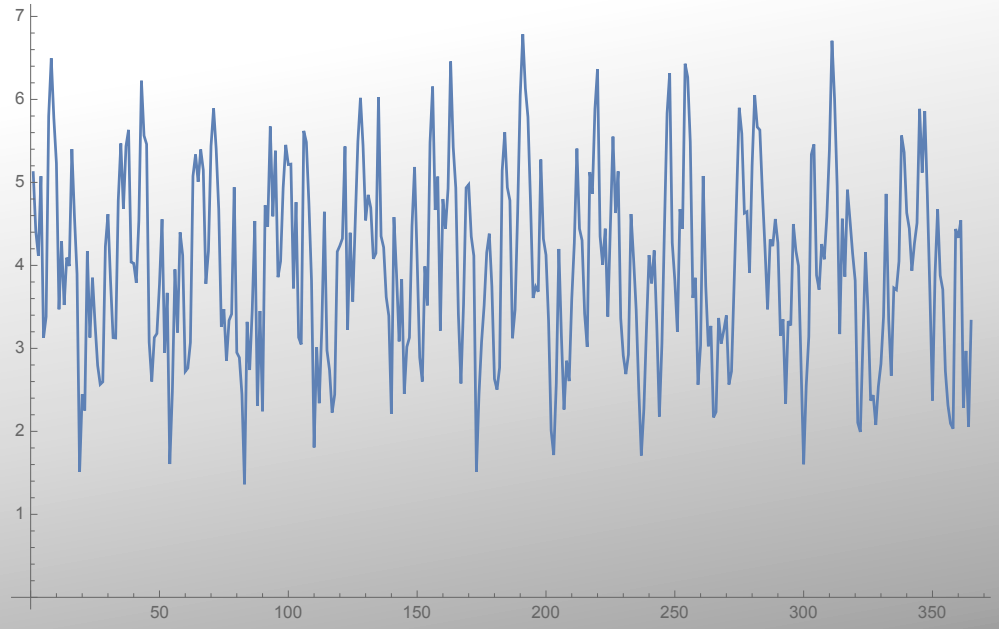
Lillian González, Rosa Garza, Sylvia Nwakanma, Mario Bañuelos, Steve Devlin, David Uminsky

MSRI- UP Team

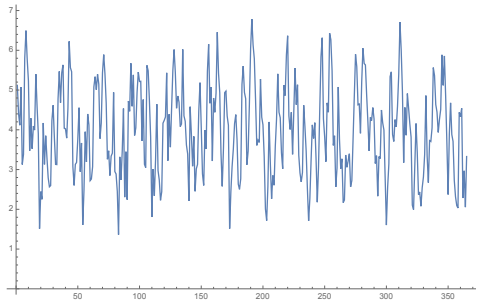
Fourier Analysis refresher

We might observe a function at times $1, 2, 3, \ldots, N$

For example, births per day for a year: $N = 365$

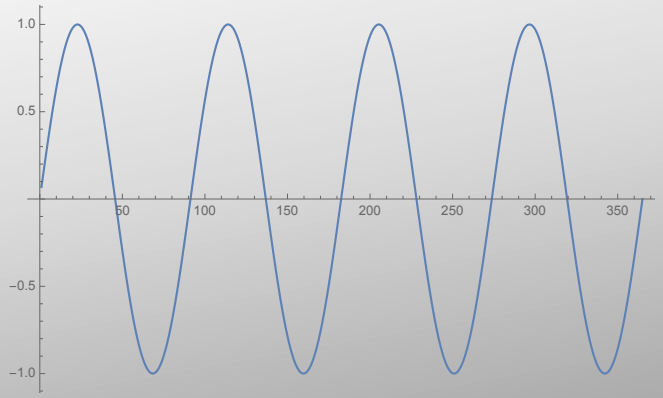


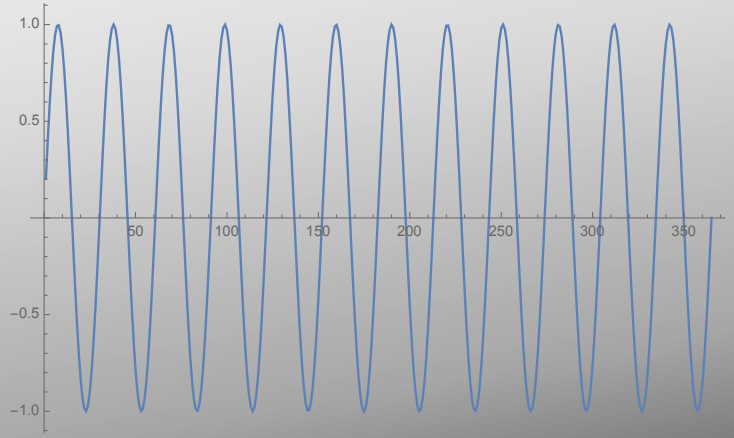
Fourier Analysis continue



Does the signal have periodicities? Weekly, Monthly, Seasonal?

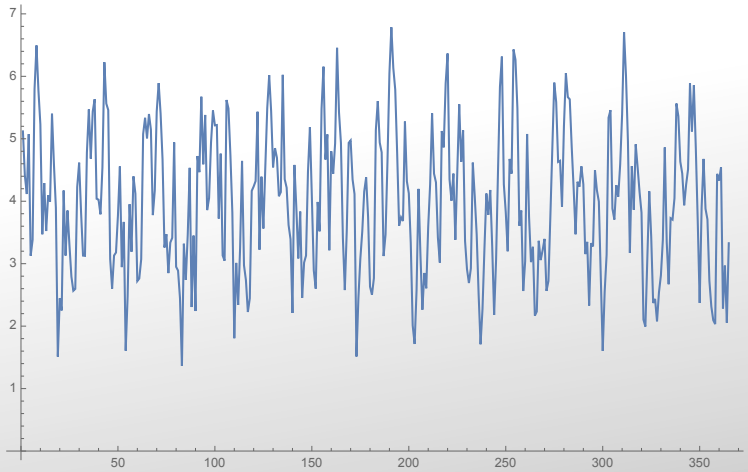
Explore the spectrum of frequencies…

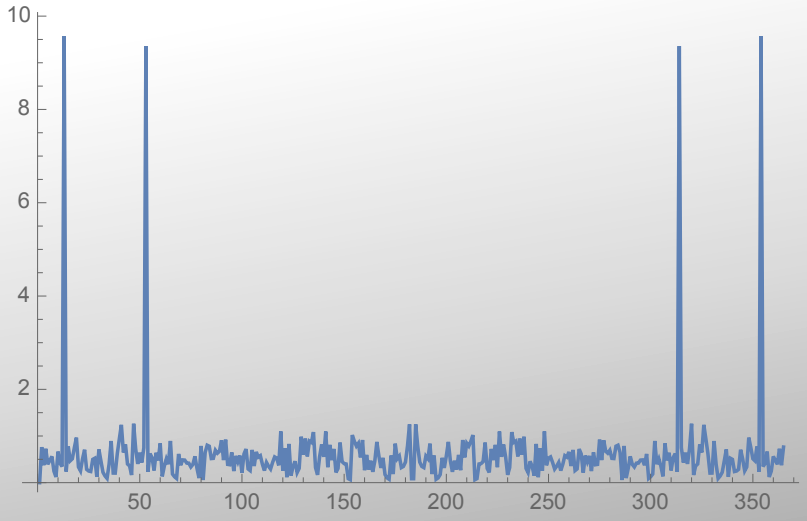




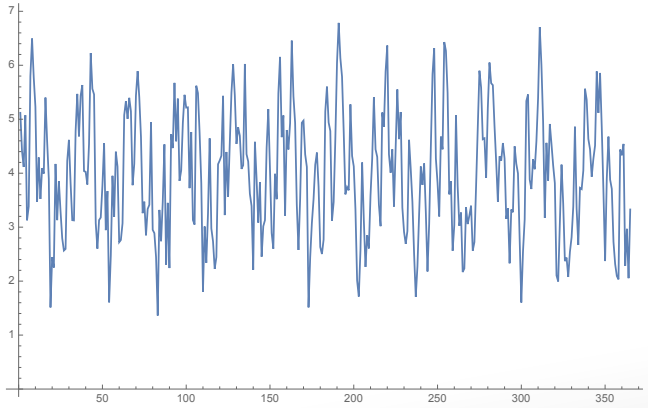
Fourier Analysis…

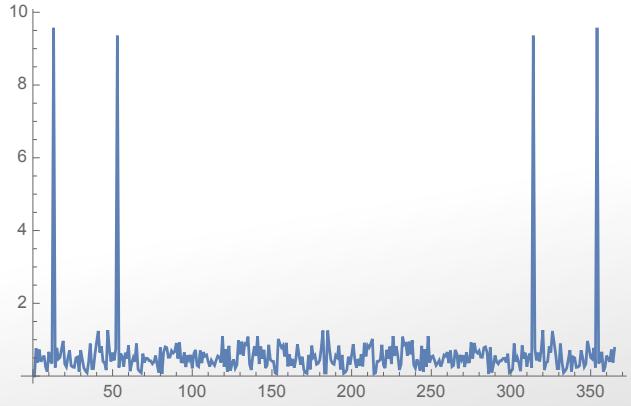
Time domain → Frequency domain





Fourier Analysis…





$$f(x) = \sum_n \langle f, \phi_n \rangle \phi_n$$

Fourier Transform (FT)

$$f(x) = \sum_n \langle f, \phi_n \rangle \phi_n$$

$f(x)$ a function on a set…

Expand f in a new basis that gives insight to the function

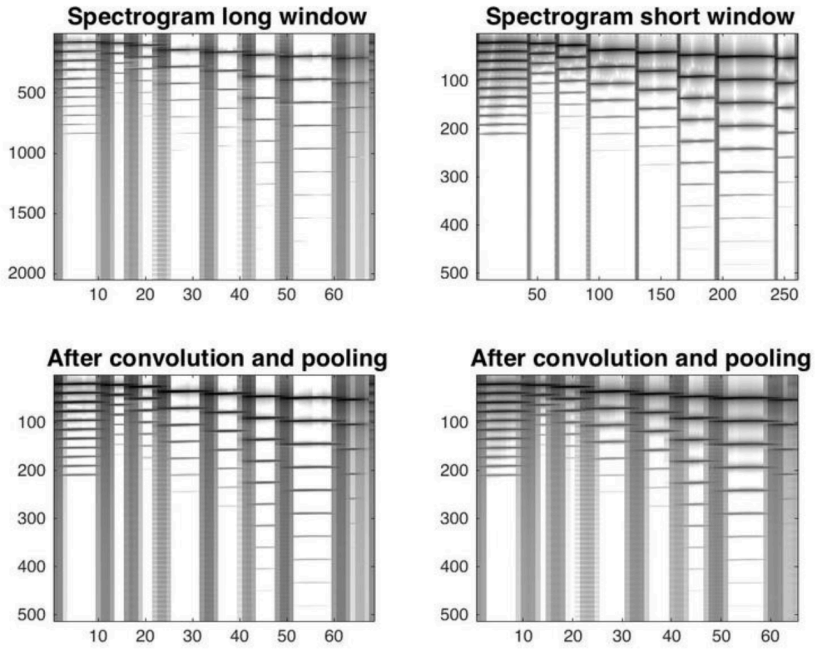
Fourier Analysis

Under the hood…

The basis functions ϕ_n are periodic functions (Sines and Cosines) on the group $\mathbb{Z}/N\mathbb{Z}$.

Connections of FT to DL/ML etc…

Ex: Audio Processing, (Doefler, SAMPTA 2017)



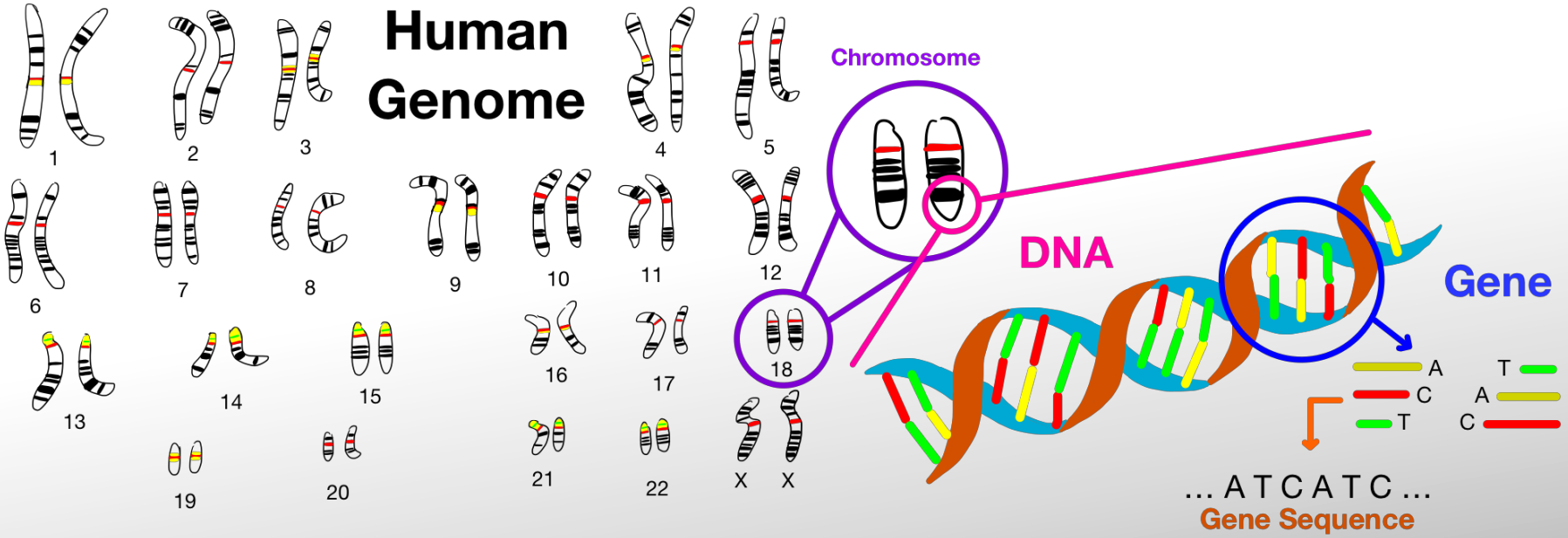
Synthetic signal consisting of several damped notes. Left upper plot shows spectrogram obtained with long window g, upper right plot shows spectrogram obtained with short window. Lower plot show output of convolution and subsequent max-pooling chosen to result in equivalent matrix size. It is visible, that the convolution leads to similar results from previously different spectrograms.

Epistasis

What is Epistasis?

- One of the major reasons why medicine isn't "solved" when humans were sequenced.
- Specifically it is the issue that it is unlikely that a single mutation causes a phenotypic response but instead it is likely a higher order interaction between a collection of genes.
- Classical computational techniques like GWAS and QTL are successful in identifying one or more genes who may contribute individually to a measured phenotype.
- Higher order interaction is a severe challenge computationally and statistically (Bonferoni corrections etc…). $\binom{N}{k}$ >>> N and N is already likely to be big.

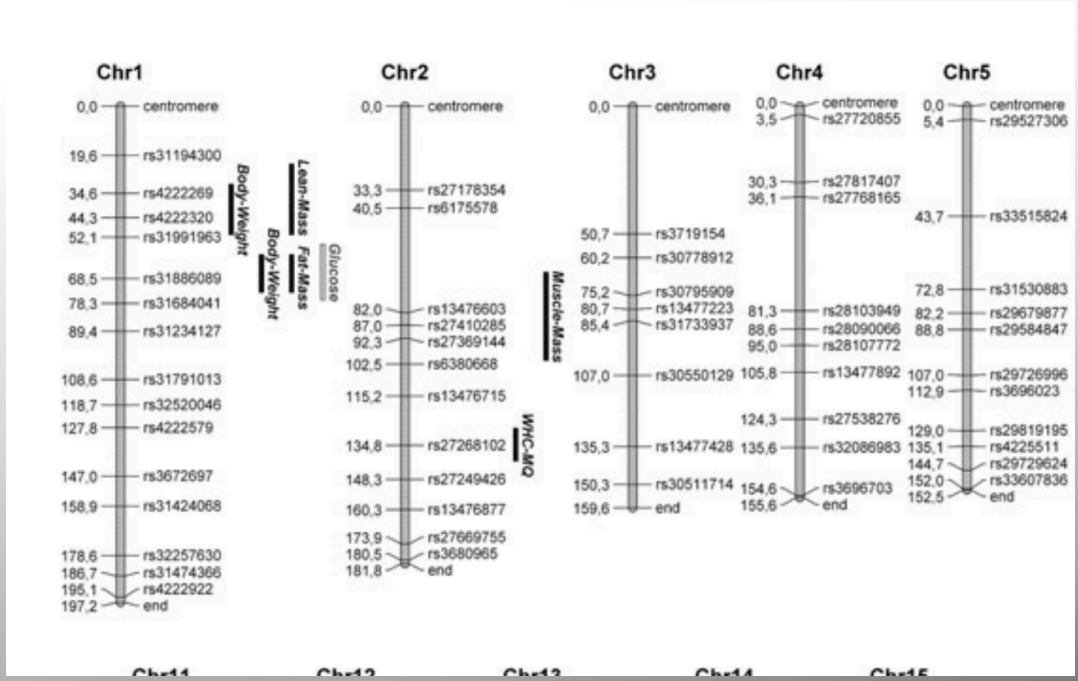
A bit more of the biology



- Human genomic data with identified, measured quantitative phenotypes are... challenging.
- Berlin Muscle Mice have less genomic privacy concerns and just safer place to play with FTs.

The Data (Karst et al)

- 332 inbred, Berlin Muscle Mice.
- 40 chromosomes, Important to consider male and female genomic data separately.
- Several phenotypes measured, we focused on body weight.
- 164 SNP identified for Quantitative Trait Loci (QTL) Analysis.



What can Fourier transforms do here?

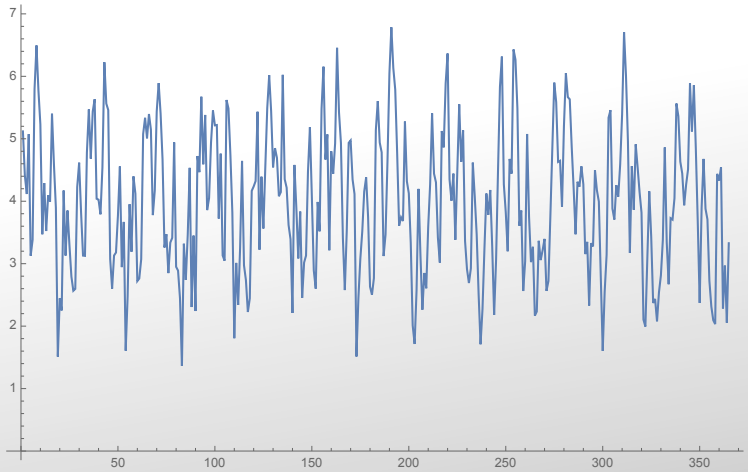
- We focus on Chromosome 1 (15 mutations, with labels A-O).
- Karst et al identify mutations, B C and E as individual positive contributor to body weight phenotype.
- Can we detect epistasis here using FT?
- What group should we use? $\mathbb{Z}/N\mathbb{Z}$ is the structure.
- Mutation data are functions on the power set of N genes. It turns out that S_N, the symmetric group on N genes has the correct group structure.
- Associated irreducible representations create a **new basis** that orthogonally decomposes gene data into an interpretable spectrum.
- Invariance ⟶ Group effects
- Orthogonality ⟶ Pair effects are removed from groups of three, four, etc.

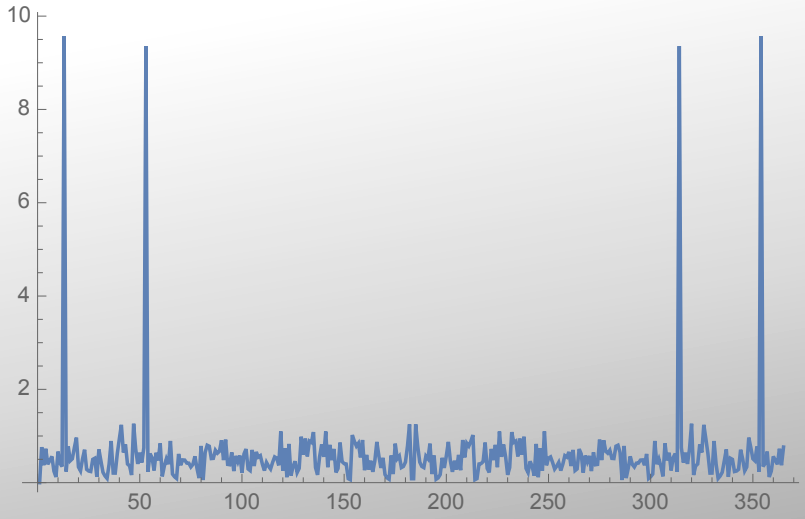
FT on Chromosome One.

- 15 identify mutation locations, label {A, B, C,….,N, O}.
- For each mouse identify which of 15 have mutation e.g. {F,N} or {B,C,D,E}
- Record mouse weight: f({F,N}) = 15, f({B,C,D,E}) = 22 ….
- Our function f is our data vector recording all body weight as a function of subset mutations.
- Compute FT against S_N on f.

Reminder…

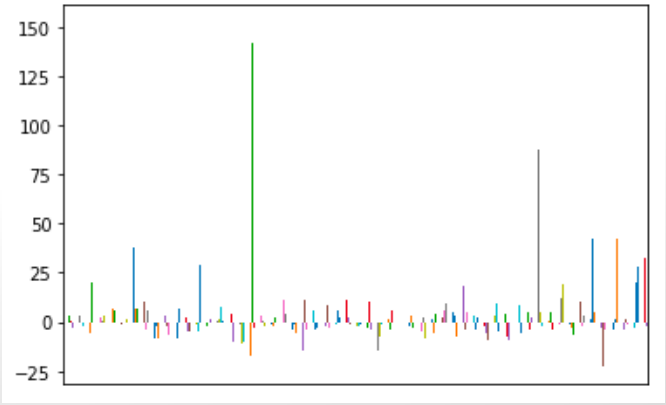
Time domain → Frequency domain



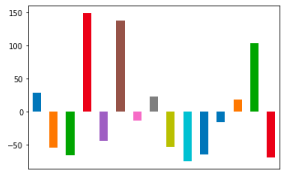


Example FT for 5 mutations

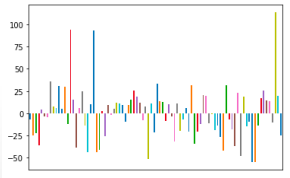
Body mass



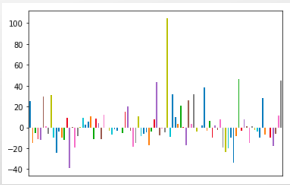
Ind.



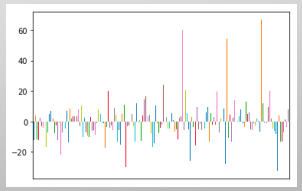
Pairs



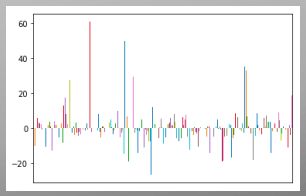
3s



4s

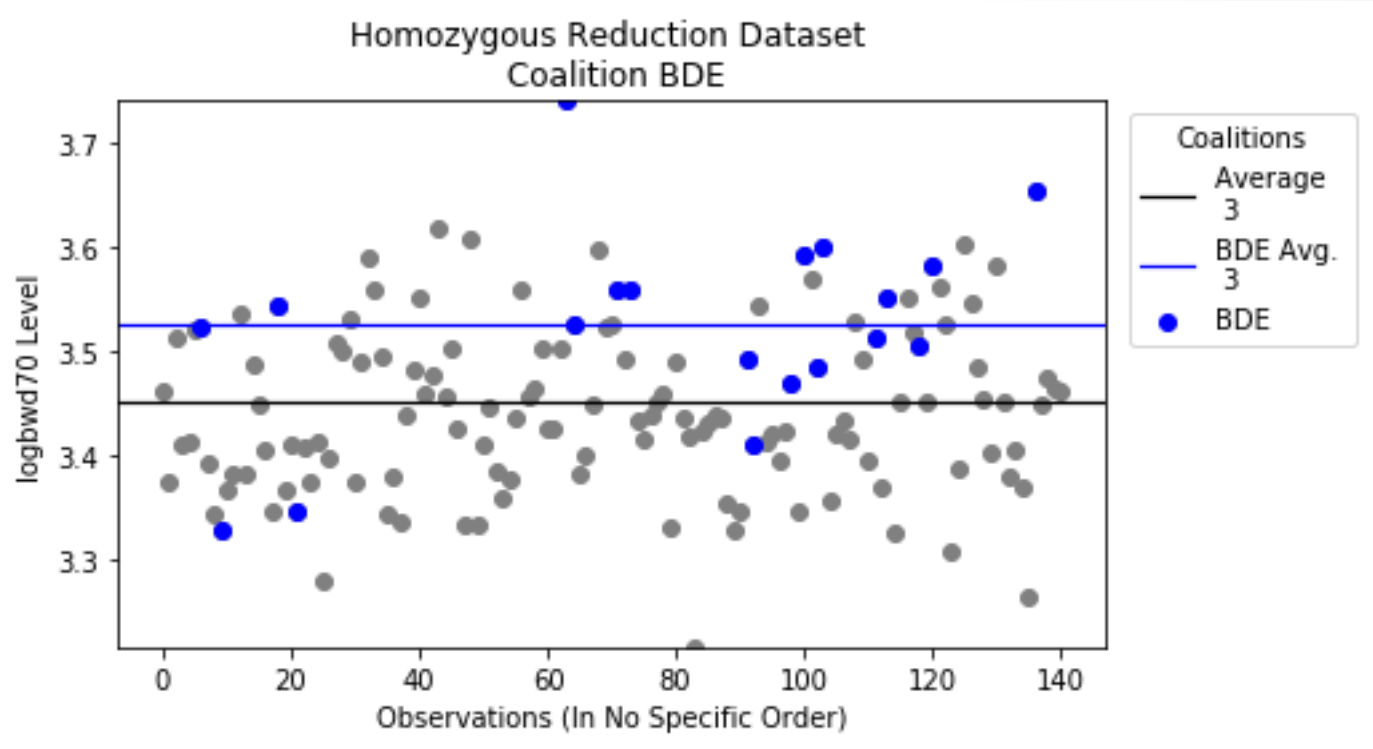


5s



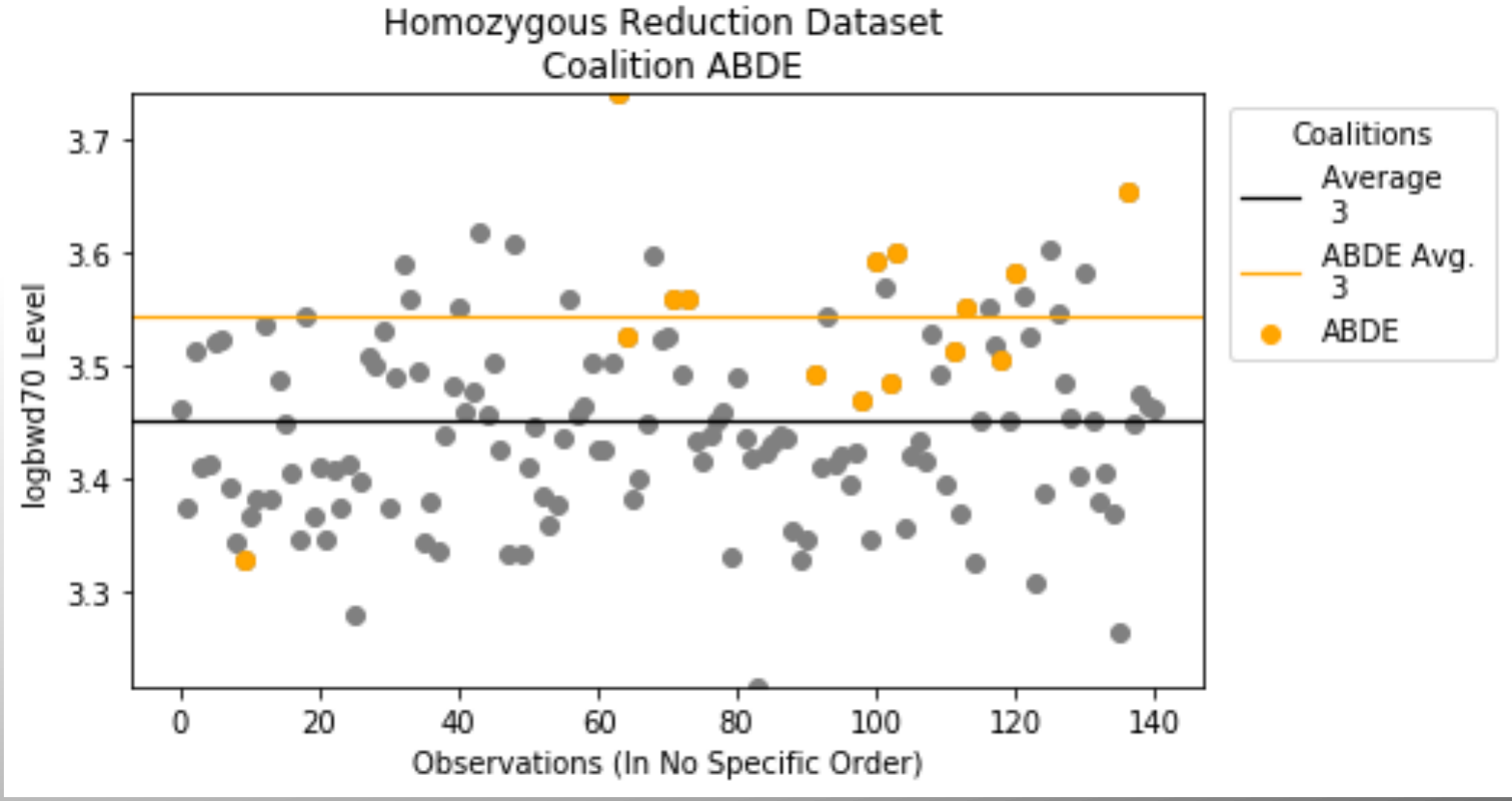
Results

- Recall that the main finding of Karst et al. is that mutations {B}, {C}, {E} are positive contributors to body weight.
- **Result #1:** Epistasis detected. Specifically {B,C,E} has a statistically signficant lift on body mass!



Results

- **Result #2:**, {A,B,C,E} also has a statistically significant lift on body mass. FT identifies {A,B,C, E} orthogonally, significant to {B,C,E}.



Conclusion

- To our knowledge, first general technique to detect statistically significant 3rd and 4th order epistasis effect. Standard QTL and GWAS identify single mutations and state of the art methods attempt to detect pair effects.

- We partially confirm Karst results that B and D are significant but not C and only if {B,D,E} are mutated together.

- We have discovered that mutation A, previously overlooked, can trigger a further increase in body weight when mutation occurs along with B, D, and E.

- Next Steps in direction of ML:
  Application to DL/ML - FT are also just convolutions.
  Model building around group structure of FT.

Thank you!

- Thank you to Laura Montoya and Program Committee of LXAI, amazing event!
- Thank you to MSRI for hosting this research.
- Finally we would like to thank our sponsors National Science Foundation
- (DMS-1659138), National Security Agency (H98230-18-1-0008), and Alfred P. Sloan Foundation (G-2017-9876).

MSRI
Mathematical Sciences Research Institute

NSF

NATIONAL SECURITY AGENCY
UNITED STATES OF AMERICA

ALFRED P. SLOAN FOUNDATION
1934
Alfred P. Sloan
FOUNDATION