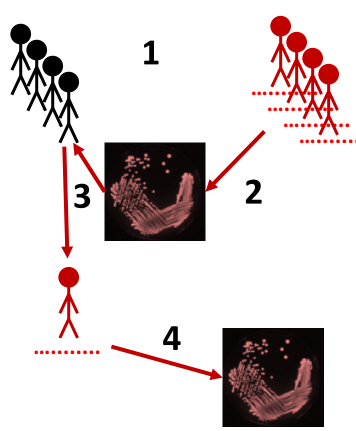
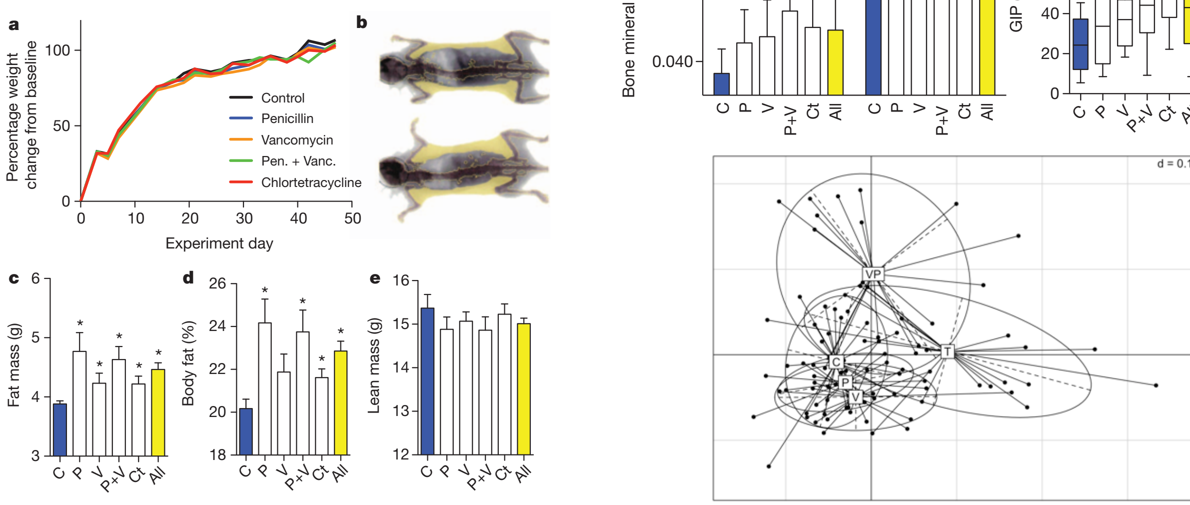
Multivariate mediation analysis with microbiome data

The role of the microbiome in human disease, an infectious disease approach

- Robert Koch's (1843 - 1910) postulates:
    1. The microorganism must be found in abundance in all organisms suffering from the disease, *but should not be found in healthy organisms*.
    2. The microorganism must be isolated from a diseased organism and grown in pure culture.
    3. The cultured microorganism should cause disease when introduced into a healthy organism.
    4. The microorganism must be reisolated from the inoculated, diseased experimental host and identified as being identical to the original specific causative agent.



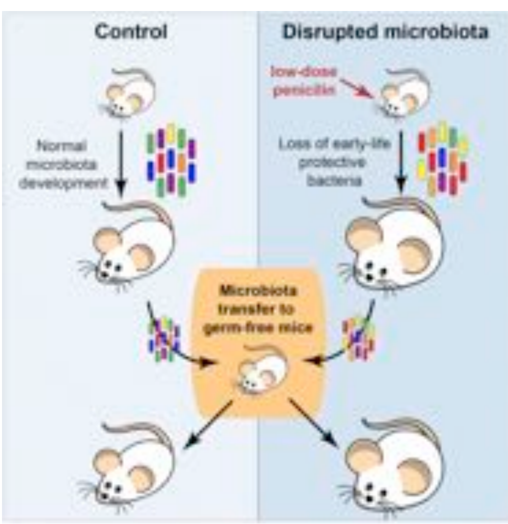


doi:10.1038/nature11400

Antibiotics in early life alter the murine colonic microbiome and adiposity

Ilseung Cho[1,2], Shingo Yamanishi[1], Laura Cox[3], Barbara A. Methé[4], Jiri Zavadil[5,6], Kelvin Li[4], Zhan Gao[3], Douglas Mahana[3], Kartik Raju[3], Isabel Teitler[3], Huilin Li[7], Alexander V. Alekseyenko[1,6] & Martin J. Blaser[1,2,3]

Causality via germ-free experiments (postulate 3)

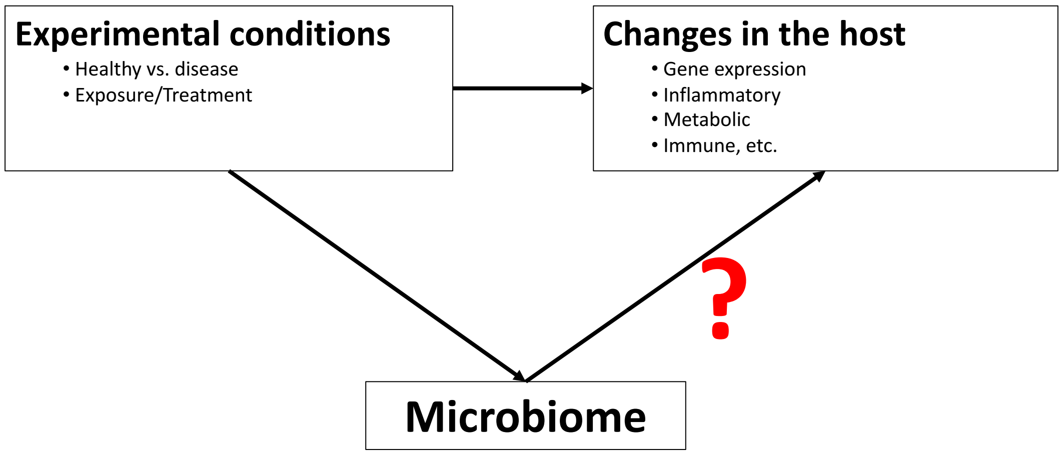


Cell

Altering the Intestinal Microbiota during a Critical Developmental Window Has Lasting Metabolic Consequences

Laura M. Cox,[1,2] Shingo Yamanishi,[2] Jiho Sohn,[2] Alexander V. Alekseyenko,[2,3] Jacqueline M. Leung,[1] Ilseung Cho,[2] Sungheon G. Kim,[4] Huilin Li,[5] Zhan Gao,[2] Douglas Mahana,[1] Jorge G. Zárate Rodriguez,[7] Arlin B. Rogers,[6] Nicolas Robine,[8] P'ng Loke,[1] and Martin J. Blaser[1,2,9,*]

Microbiome as a mediator in human health



Single Mediator Model (SMM)

Let

- X be independent,
- Y be dependent, and
- M be mediator

variables SMM models linear relationships between them as

- $Y = i_1 + cX + e_1$;
- $Y = i_2 + c'X + bM + e_2$;
- $M = i_3 + aX + e_3$,

- where
  - i_1, i_2, i_3 are intercepts,
  - c and c' quantify unadjusted and adjusted effect of the independent variable on the dependent one, and
  - a is relating independent variable and the mediator,
  - b is relating the mediator and the dependent variable, and
  - e_1, e_2, e_3 model unexplained variability.

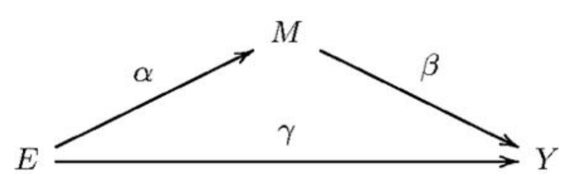
Multiple Mediator Models for Microbiome Data

- Jie Zhang, Zhi Wei and Jun Chen “A distance-based approach for testing the mediation effect of the human microbiome” Bioinformatics 2018
  - Joint test of conditional association of the mediator with response, based on eigen decomposition of the multivariate distance metrics.
- Michael B. Sohn and Hongzhe Li “Compositional Mediation Analysis for Microbiome Studies” bioArxiv
  - Takes into account compositional nature of microbiome data
  - Builds a shrinkage-based estimation procedure to establish mediation
- All of these allow for univariate exposure and response and multivariate mediator

Testing framework for mediation analysis

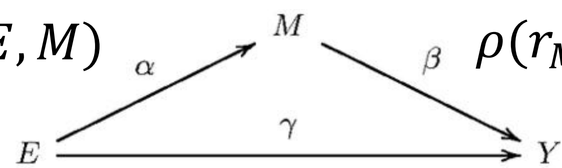
Simina M. Boca, Rashmi Sinha, Amanda J. Cross, Steven C. Moore, Joshua N. Sampson “Testing multiple biological mediators simultaneously”. Bioinformatics. 2014 Jan 15;30(2):214-20.

- Permutation-based test that jointly establishes significance of:
  - $\rho(E, Y)$, $\rho(E, M)$, and $\rho(r_{M|E}, r_{Y|E})$
  - Where
    - $\rho(.,.)$ is correlation
    - $r_{M|E}$ and $r_{Y|E}$ are residuals after regression of E on M and Y, respectively.



What if M is multivariate?

$\rho(E, M)$ $\rho(r_{M|E}, r_{Y|E})$



$\rho(E, Y)$

From unpublished book "Statisticians talking about fruit"

If we want to study the dependence between oranges and apples then it is hard to add or multiply them but it is always easy to do the same with their distances.

- Gábor J. Székely

… current techniques of meta-analysis do little more than take weighted averages of the various studies, thus averaging apples and oranges to infer properties of bananas.

- Judea Pearl

Multivariate dependence via distance correlation

Pearson correlation

- Data: $(X_k, Y_k), k = 1, \ldots, n$
- Centered data:
  - $A_k = X_k - X_.; B_k = Y_k - Y_.$
- Covariance:
  - $Cov(X,Y) = \frac{1}{n}\sum_k A_k B_k$
- Correlation:
  - $\rho(x,y) = \frac{Cov(x,y)}{\sqrt{Cov(x,x)Cov(y,y)}}$

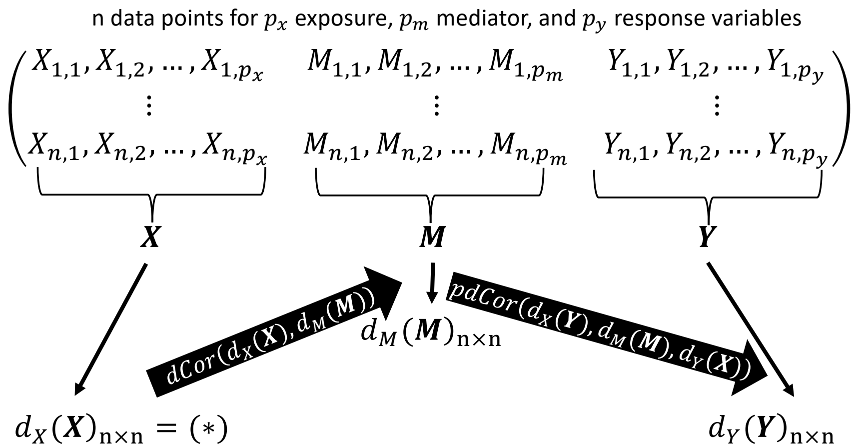
Distance correlation

- Data: $(\boldsymbol{X_k}, \boldsymbol{Y_k}), k = 1, \ldots, n$
- **Distance**:
  - $a_{kl} = |\boldsymbol{X_k} - \boldsymbol{X_l}|$;
  - $b_{kl} = |\boldsymbol{Y_k} - \boldsymbol{Y_l}|$
- Centered distance:
  - $A_{kl} = a_{kl} - a_{k.} - a_{.l} + a_{..}$;
  - $B_{kl} = b_{kl} - b_{k.} - b_{.l} + a_{..}$
- Distance covariance:
  - $dCov(X,Y) = \frac{1}{n}\sum_{kl} A_{kl} B_{kl}$

See Szekely, Rizzo, Bakirov(2007) Ann. Statist. 35/7

Partial Distance Correlation

- Let R*(X,Y) be bias corrected dCor(X,Y) (no time to develop in this talk)
- Define pdCor similarly to Pearson/Fisher partial correlation:
  - $pdCor(X,Y;Z) = \frac{R^*(X,Y) - R^*(X,Z)R^*(Y,Z)}{\sqrt{1-R^*(X,Z)^2}\sqrt{1-R^*(X,Y)^2}}$
- pdCor is valid for all X, Y, Z in arbitrary dimensions, not necessarily the same.

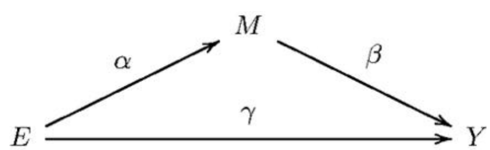


$$(*) = \begin{pmatrix} d_X(X_{1,\cdot}, X_{1,\cdot}) = 0 & d_X(X_{1,\cdot}, X_{2,\cdot}) & \cdots & d_X(X_{1,\cdot}, X_{n,\cdot}) \\ d_X(X_{2,\cdot}, X_{1,\cdot}) & \ddots & & d_X(X_{2,\cdot}, X_{n,\cdot}) \\ \vdots & & & \vdots \\ d_X(X_{n,\cdot}, X_{1,\cdot}) & d_X(X_{n,\cdot}, X_{2,\cdot}) & \cdots & d_X(X_{n,\cdot}, X_{n,\cdot}) = 0 \end{pmatrix}$$

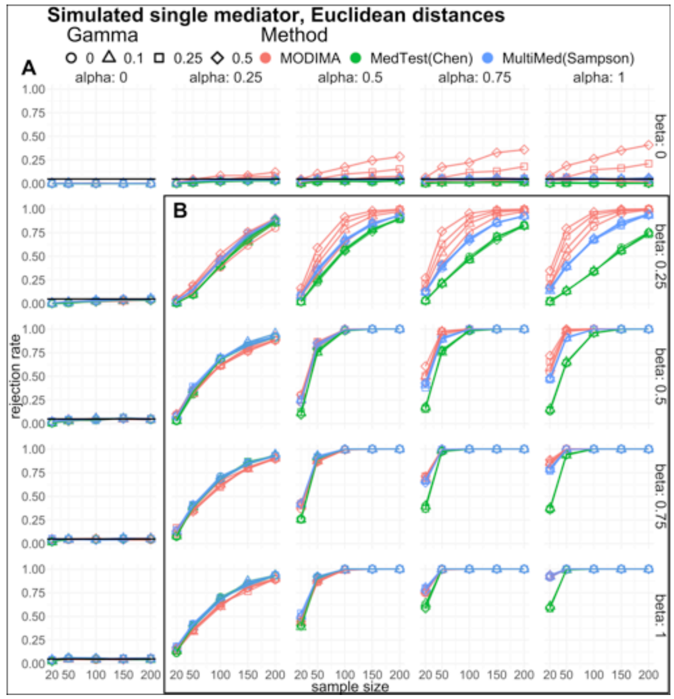
MedTest by Jie Zhang, Zhi Wei, Jun Chen

- Consider diagonalization of a doubly-centered distance matrix:
  - Eigenvectors, eigenvalues:$(\boldsymbol{u_1}, \boldsymbol{u_2}, \dots, \boldsymbol{u_l})$; $(\lambda_1, \lambda_2, \dots, \lambda_l)$
- The basis of the MedTest statistic is the weighted average of the component—wise effects:
  - $T = \sum_{l=1}^{L} \lambda_l \mid < X, \boldsymbol{u_l} > \times < Y, \boldsymbol{u_l} > \mid$
- To test for association permutations of X, Y and both are considered.
  - These disrupt the possible relationships between X and M, Y and M, or both, providing a way to estimate all three null hypotheses jointly, by taking the maximum.

Simulation results



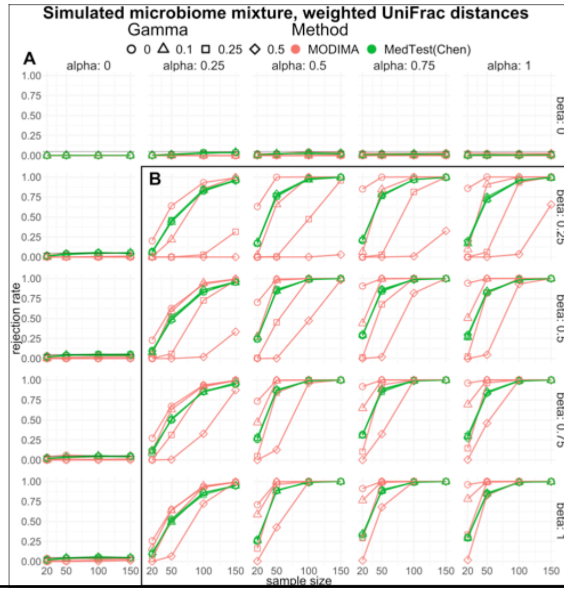
- $E \sim N(0,1)$
- $M \sim \alpha E + N(0, \sigma_M^2)$
  - $\sigma_M^2 = 1 - \alpha^2$
- $Y \sim \gamma E + \beta M + N(0, \sigma_Y^2)$
  - $\sigma_Y^2 = 1 - \sigma_M^2 * \beta^2 - (\alpha\beta + \gamma)^2$



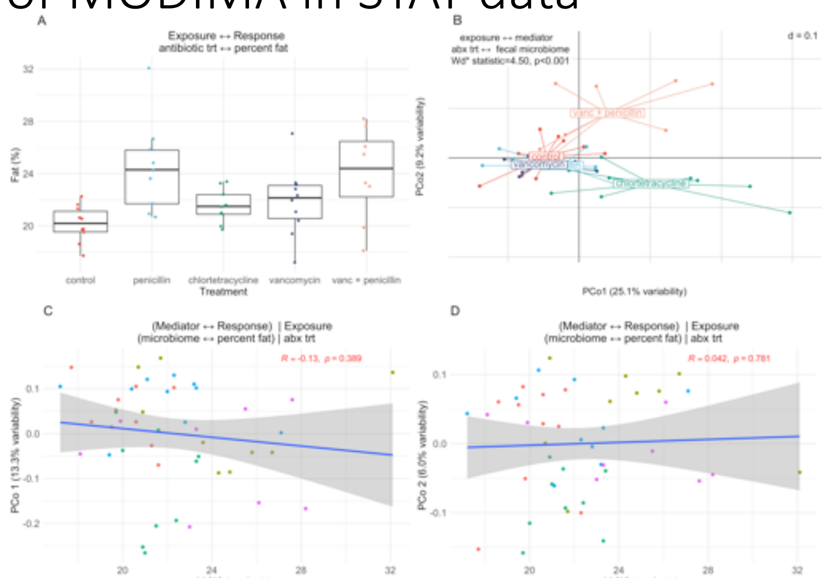
Simulated microbiome data

- Simulation
  - Draw E
  - E defines mixture coefficient alpha for two different microbiota types
  - Draw M as the mixture
  - Compute alpha diversity of M and use that as a link to mediated effect
  - Draw R from E and computed alpha diversity above

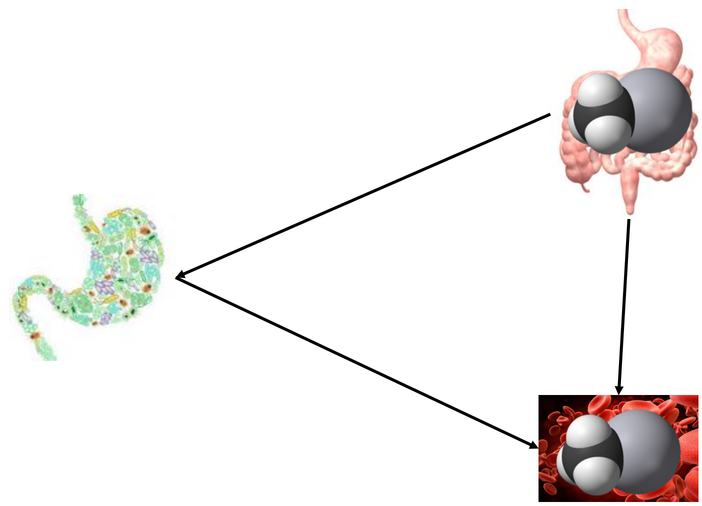
Performance in simulated microbiome data



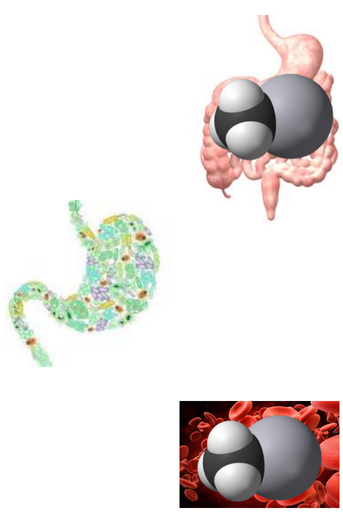
Application of MODIMA in STAT data



Application: Gut microbiome mediation of mercury absorption in pregnant women



Contemporaneous mediation



Early Pregnancy

Stool-MeHg

Dcor = 0.14

Dcor = 0.39

µBiome

Dcor = 0.10

Blood-MeHg

P = 0.21
Pdcor = 0.05
E = 0.006

Late Pregnancy

Stool-MeHg

Dcor = 0.02

Dcor = 0.05

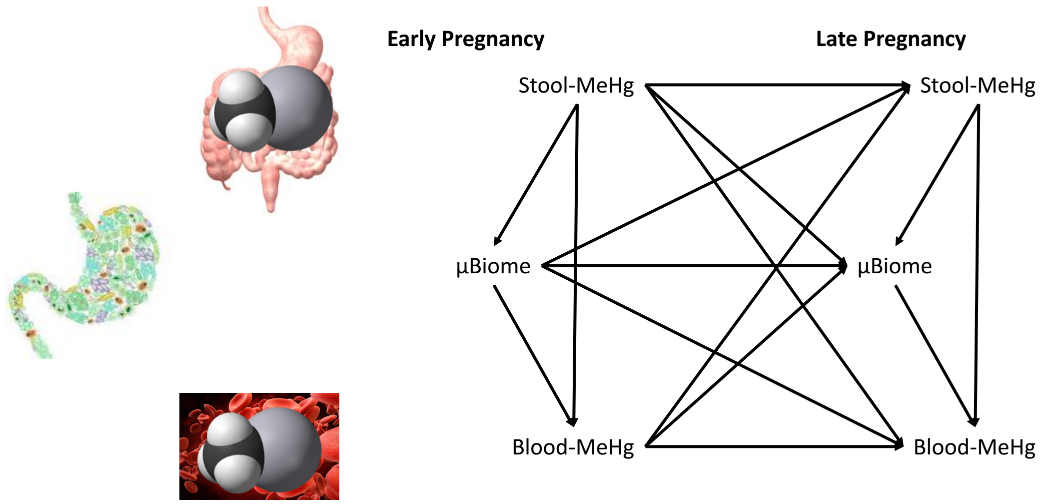
µBiome

Dcor = 0.02

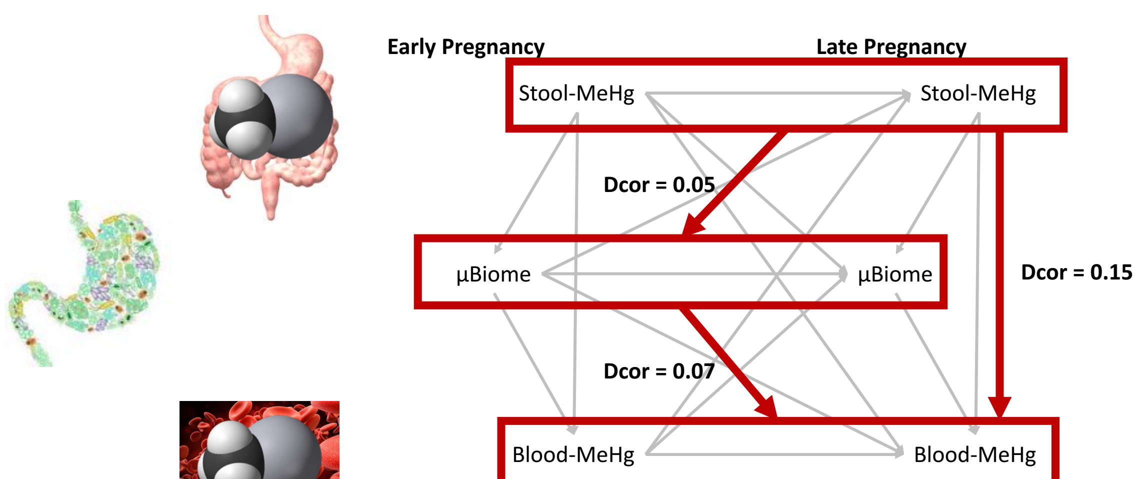
Blood-MeHg

P = 0.35
Pdcor = 0.02
E = 0.0003

The universe of all other possible causal models



Joint Mediation



P = 0.22
Pdcor = 0.06
E = 0.003

Alternative approach

- Michael B. Sohn and Hongzhe Li "Compositional Mediation Analysis for Microbiome Studies"
  - Takes into account compositional nature of microbiome data
  - Builds a shrinkage-based estimation procedure to establish mediation
  - Missing publicly available implementation