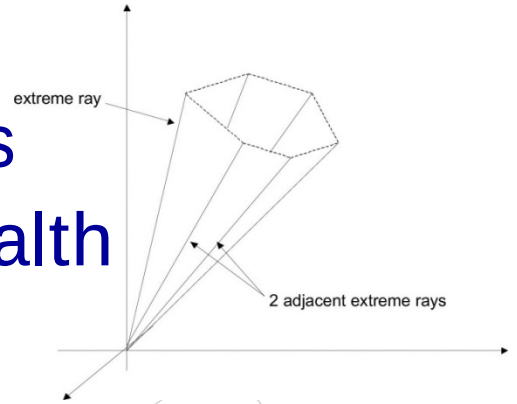
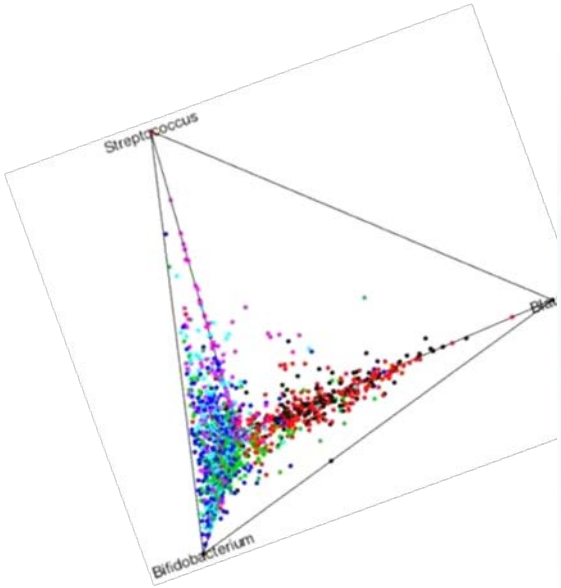
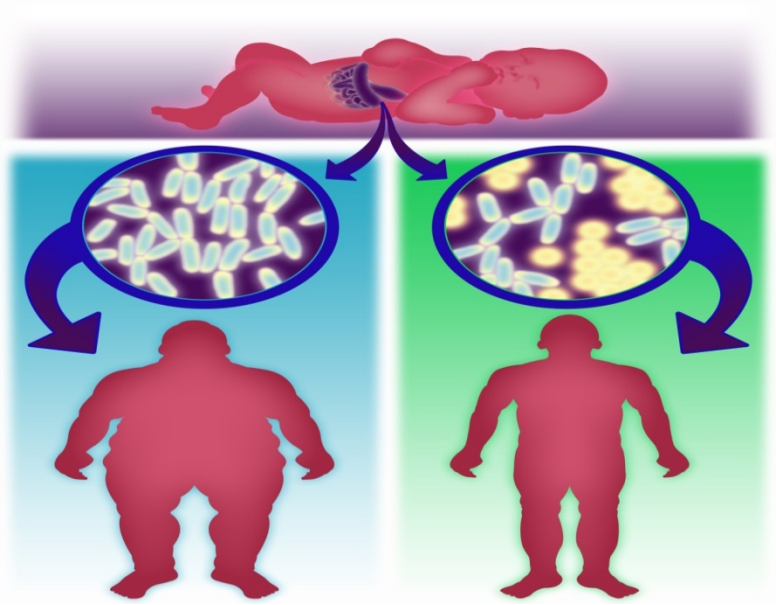
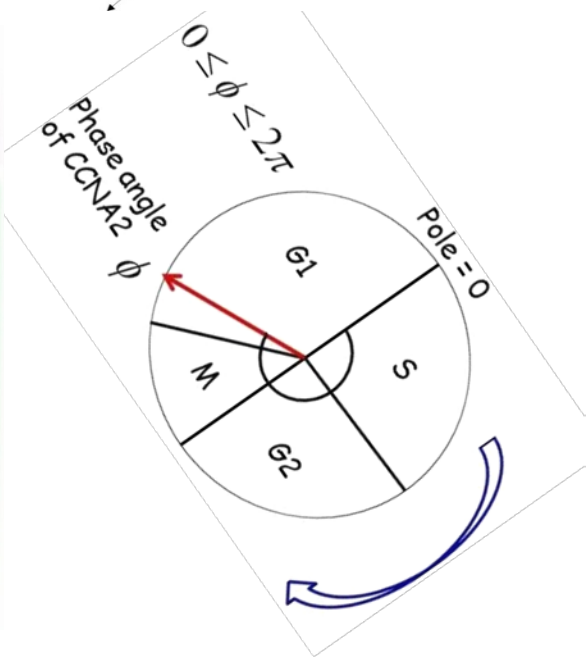
Constrained statistical inference for the analysis of microbiome data

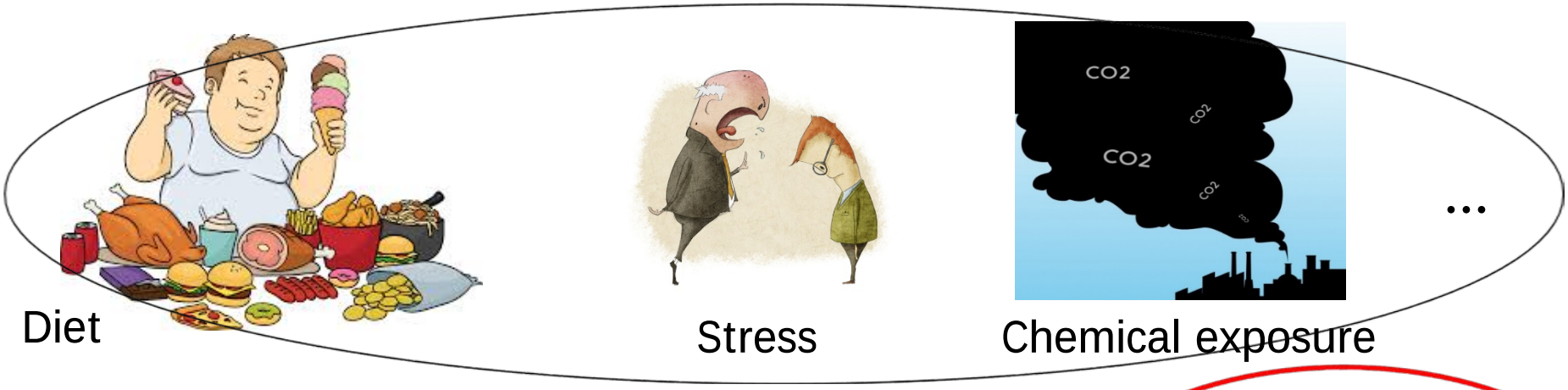
Shyamal Peddada
Department of Biostatistics
Graduate School of Public Health
University of Pittsburgh











External environment

External environment

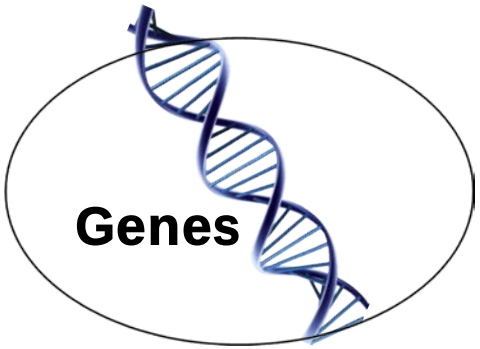
X

Internal environment

X

Genes

Time



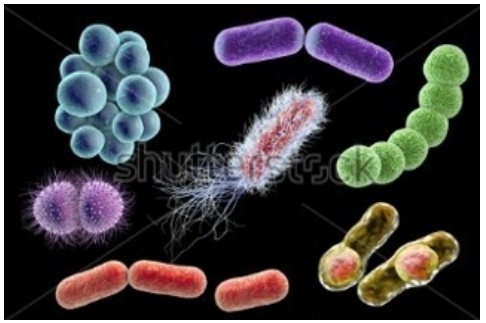
However, we are mostly microbes!

Cells:

- ~10 trillion human cells
- ~100 trillion microbial cells

Genes:

- ~20,000 human genes
- ~2 to 20 million microbial genes



Internal environment: Microbiome

Focus of today's talk

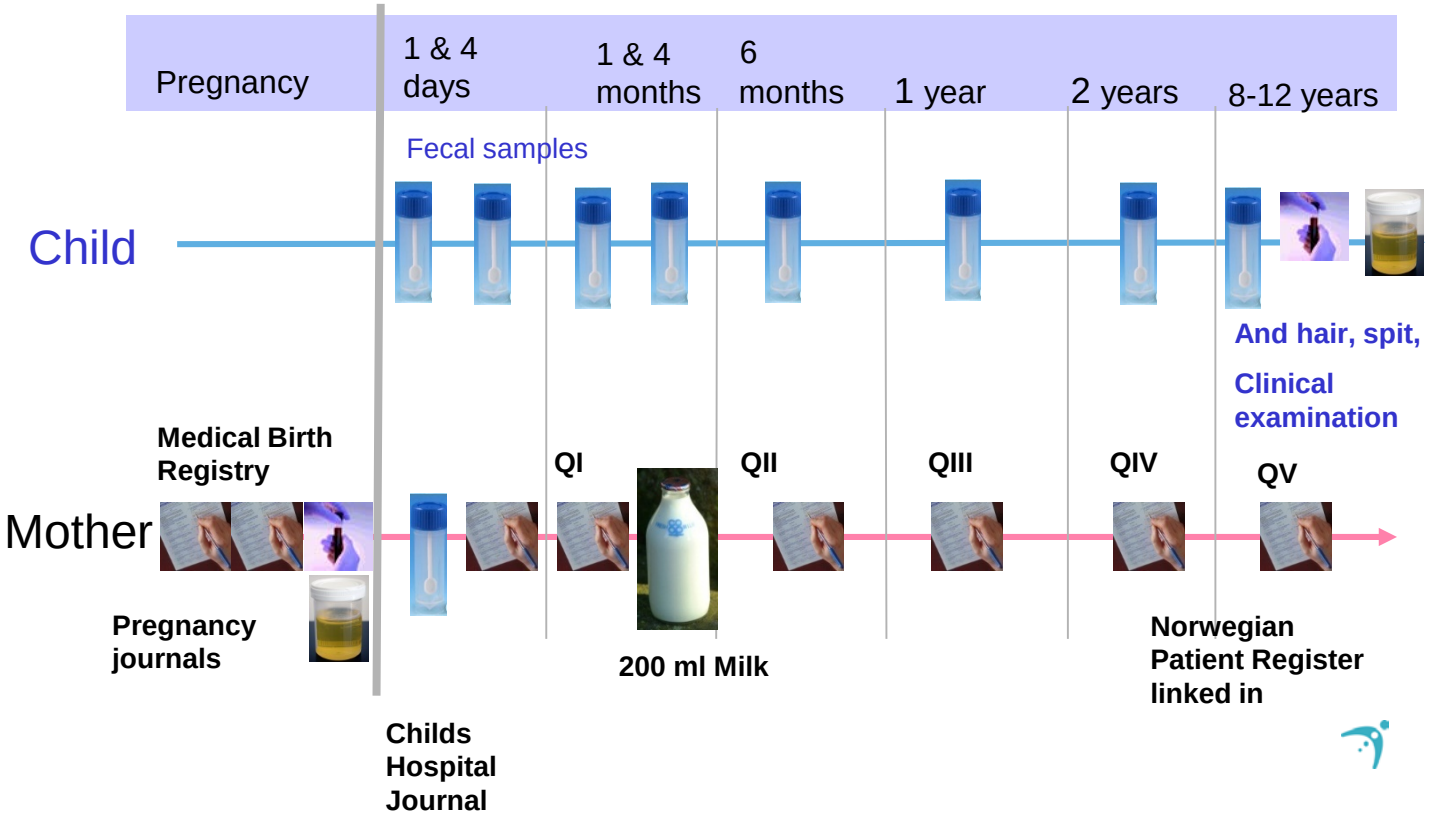
- Motivating example – Norwegian Microbiome (NoMIC) Study

- Differential abundance analysis

  - The methodology
  - Illustration: Effect of external environment factors on infant gut microbiota ...

NoMIC study of 550 infants

PI Merete Eggesbø, NIPH, Oslo, Norway



Data

Ecosystem (e.g. gut):

↓

A random specimen

- Sequence the specimen
- Read counts of 16S rRNA for each Operational Taxonomic Unit (OTU)

OTU Abundance Table

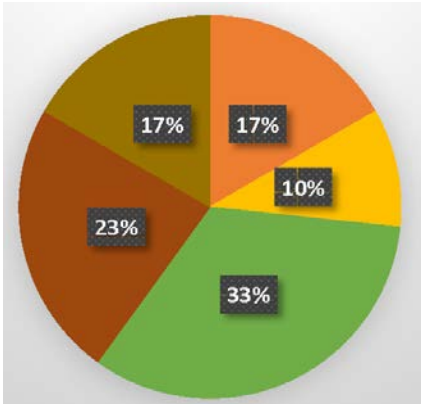
OTU	Subject 1	Subject 2	…	Subject n
OTU_1	O_{11}	O_{12}	…	O_{1n}
OTU_2	O_{21}	O_{22}	…	O_{2n}
OTU_3	O_{31}	O_{32}	…	O_{3n}
OTU_4	O_{41}	O_{42}	…	O_{4n}
…	…	…	…	
OTU_m	O_{m1}	O_{m2}	…	O_{mn}

A tale of two types of parameters …

Abundance Vs. Relative Abundance

Abundance of 5 taxa: Ecosystem

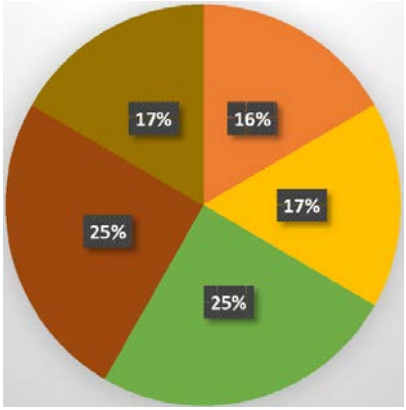
Relative abundance of 5 taxa: Ecosystem



Unobservable

Abundance of 5 taxa: Specimen

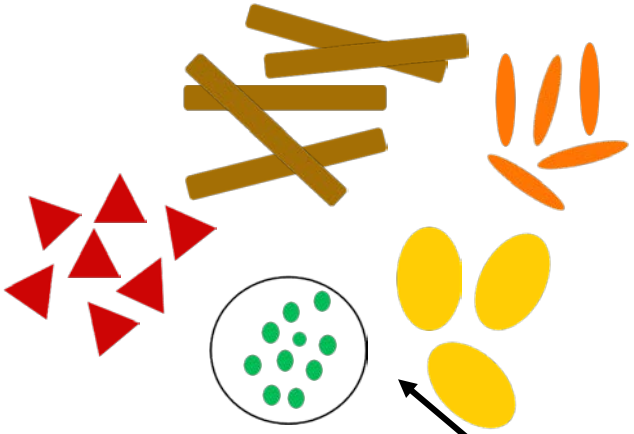
Relative abundance of 5 taxa: Specimen



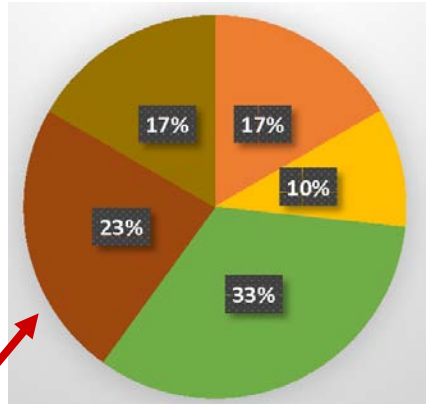
Observable

A Single Taxon Can Change all Relative Abundances

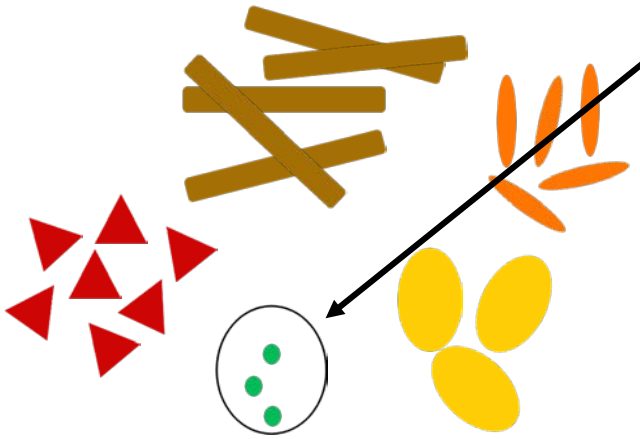
Abundance of 5 taxa: Ecosystem I



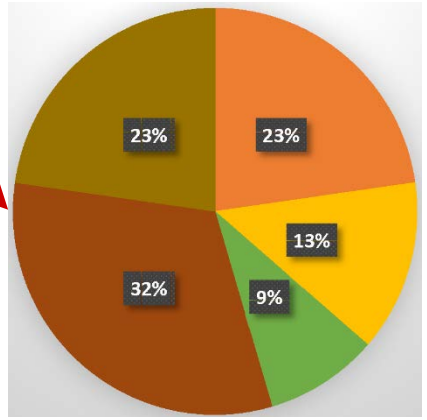
Relative abundance of 5 taxa: Ecosystem I



Abundance of 5 taxa: Ecosystem II



Relative abundance of 5 taxa: Ecosystem II



Not Sufficient to Compare Relative Abundances

Researcher may be interested in identifying taxa whose abundance changed between the ecosystems even though true abundances cannot be estimated!

Differential abundance of taxa in two or more ecosystems …

Analysis of Composition of Microbiomes (ANCOM) ...

Basic idea ...

Lemma

For $i = 1,2...,m,$ let $E(\ln(\mu_i^1) - \ln(\mu_i^2)) = d_i$

Assumption: Among $d_1, d_2, ..., d_m$ at least 2 are zero [i.e. abundance of at least 2 taxa does not change]

Lemma: Suppose for a taxon j

$$E(\ln(\mu_j^1) - \ln(\mu_r^1)) \neq E(\ln(\mu_j^2) - \ln(\mu_r^2))$$

Relative abundance

for all $r \neq j$

Then

$$E(\ln(\mu_j^1)) \neq E(\ln(\mu_j^2))$$

Abundance

Relative Abundance Data Can Be Used to Infer About Abundance: Illustration of a Lemma

Abundance Table

Taxon	Ecosystem 1	Ecosystem 2
Taxon1	1	1
Taxon2	4	4
Taxon3	10	10
Taxon4	20	100
Taxon5	65	85
Sum	100	200

Relative Abundance Table

Taxon	Ecosystem 1	Ecosystem 2
Taxon1	.01	.005
Taxon2	.04	.02
Taxon3	.10	.05
Taxon4	.20	.5
Taxon5	.65	.425

Log Relative Abundance Ratios

Taxon	Ecosystem 1	Ecosystem 2
Log(Taxon1/Taxon2)	-1.39	-1.39
Log(Taxon1/Taxon3)	-2.3	-2.3
Log(Taxon1/Taxon4)	-3	-4.61
Log(Taxon1/Taxon5)	-4.17	-4.44

$$W_1 = \#\{\text{Distinct log-ratios}\} = 2$$

Relative Abundance Data Can Be Used to Infer About Abundance: Illustration of a Lemma

Abundance Table

Taxon	Ecosystem 1	Ecosystem 2
Taxon1	1	1
Taxon2	4	4
Taxon3	10	10
Taxon4	20	100
Taxon5	65	85
Sum	100	200

Relative Abundance Table

Taxon	Ecosystem 1	Ecosystem 2
Taxon1	.01	.005
Taxon2	.04	.02
Taxon3	.10	.05
Taxon4	.20	.5
Taxon5	.65	.425

Log Relative Abundance Ratios

Taxon	Ecosystem 1	Ecosystem 2
Log(Taxon2/Taxon1)	1.39	1.39
Log(Taxon2/Taxon3)	-0.92	-0.92
Log(Taxon2/Taxon4)	-1.61	-3.22
Log(Taxon2/Taxon5)	-2.79	-3.06

$$W_1 = \#\{\text{Distinct log-ratios}\} = 2$$

$$W_2 = \#\{\text{Distinct log-ratios}\} = 2$$

Relative Abundance Data Can Be Used to Infer About Abundance: Illustration of a Lemma

Abundance Table

Taxon	Ecosystem 1	Ecosystem 2
Taxon1	1	1
Taxon2	4	4
Taxon3	10	10
Taxon4	20	100
Taxon5	65	85
Sum	100	200

Relative Abundance Table

Taxon	Ecosystem 1	Ecosystem 2
Taxon1	.01	.005
Taxon2	.04	.02
Taxon3	.10	.05
Taxon4	.20	.5
Taxon5	.65	.425

Log Relative Abundance Ratios

Taxon	Ecosystem 1	Ecosystem 2
Log(Taxon3/Taxon1)	2.30	2.30
Log(Taxon3/Taxon2)	0.92	0.92
Log(Taxon3/Taxon4)	-0.69	-2.30
Log(Taxon3/Taxon5)	-1.87	-2.14

$$W_1 = \#\{\text{Distinct log-ratios}\} = 2$$

$$W_2 = \#\{\text{Distinct log-ratios}\} = 2$$

$$W_3 = \#\{\text{Distinct log-ratios}\} = 2$$

Relative Abundance Data Can Be Used to Infer About Abundance: Illustration of a Lemma

Abundance Table

Taxon	Ecosystem 1	Ecosystem 2
Taxon1	1	1
Taxon2	4	4
Taxon3	10	10
Taxon4	20	100
Taxon5	65	85
Sum	100	200

Relative Abundance Table

Taxon	Ecosystem 1	Ecosystem 2
Taxon1	.01	.005
Taxon2	.04	.02
Taxon3	.10	.05
Taxon4	.20	.5
Taxon5	.65	.425

Log Relative Abundance Ratios

Taxon	Ecosystem 1	Ecosystem 2
Log(Taxon4/Taxon1)	3.00	4.61
Log(Taxon4/Taxon2)	1.61	3.22
Log(Taxon4/Taxon3)	0.69	2.30
Log(Taxon4/Taxon5)	-1.18	0.16

$$W_1 = \#\{\text{Distinct log-ratios}\} = 2$$

$$W_2 = \#\{\text{Distinct log-ratios}\} = 2$$

$$W_3 = \#\{\text{Distinct log-ratios}\} = 2$$

$$W_4 = \#\{\text{Distinct log-ratios}\} = 4$$

Relative Abundance Data Can Be Used to Infer About Abundance: Illustration of a Lemma

Abundance Table

Taxon	Ecosystem 1	Ecosystem 2
Taxon1	1	1
Taxon2	4	4
Taxon3	10	10
Taxon4	20	100
Taxon5	65	85
Sum	100	200

Relative Abundance Table

Taxon	Ecosystem 1	Ecosystem 2
Taxon1	.01	.005
Taxon2	.04	.02
Taxon3	.10	.05
Taxon4	.20	.5
Taxon5	.65	.425

Log Relative Abundance Ratios

Taxon	Ecosystem 1	Ecosystem 2
Log(Taxon5/Taxon1)	4.17	4.44
Log(Taxon5/Taxon2)	2.79	3.06
Log(Taxon5/Taxon3)	1.87	2.14
Log(Taxon5/Taxon4)	1.18	-0.16

$$W_1 = \#\{\text{Distinct log-ratios}\} = 2$$

$$W_2 = \#\{\text{Distinct log-ratios}\} = 2$$

$$W_3 = \#\{\text{Distinct log-ratios}\} = 2$$

$$W_4 = \#\{\text{Distinct log-ratios}\} = 4$$

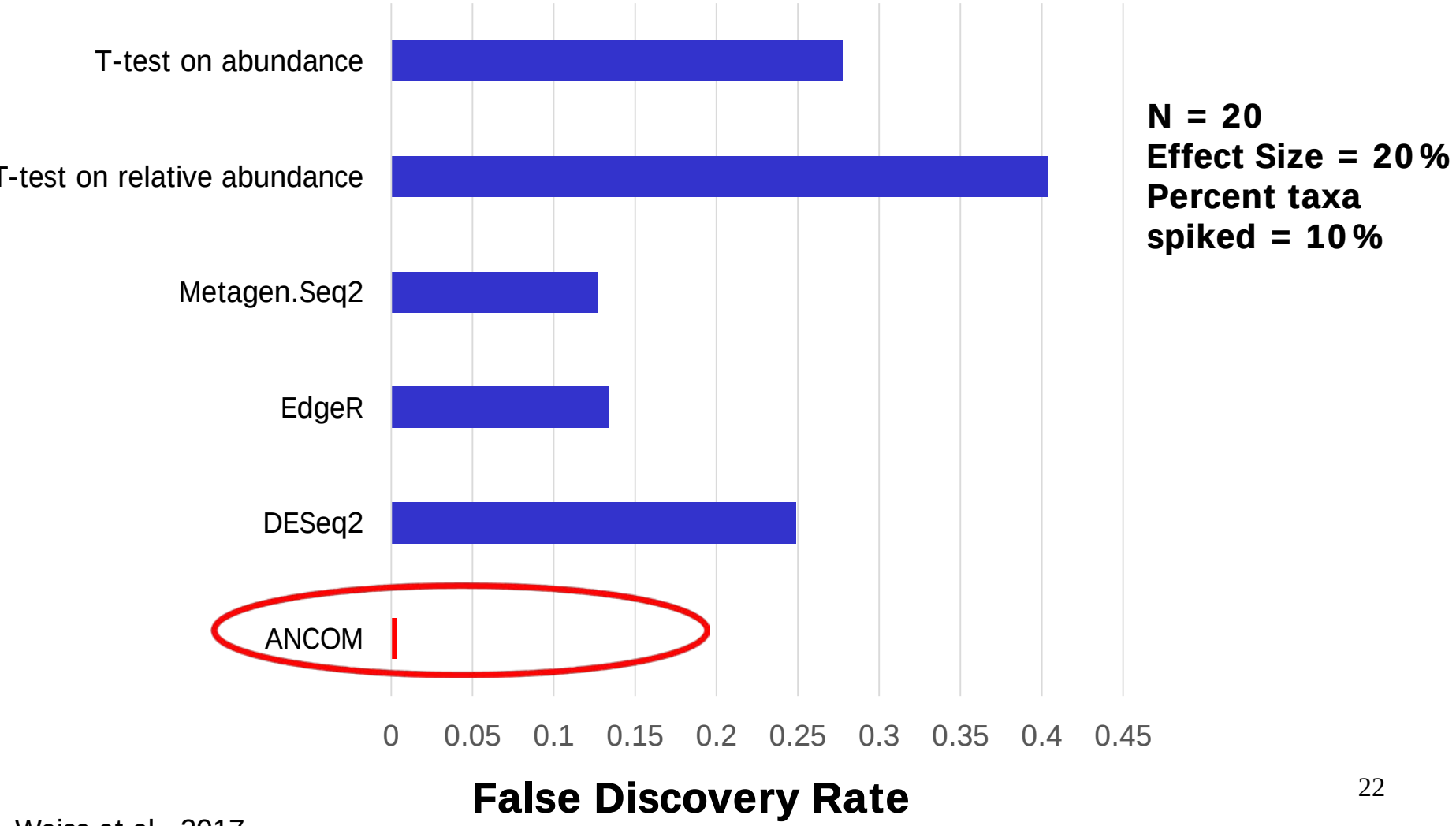
$$W_5 = \#\{\text{Distinct log-ratios}\} = 4$$

A simulation study ...

Simulation Study Based on a Real Data Set in Caporaso et al., PNAS 2011

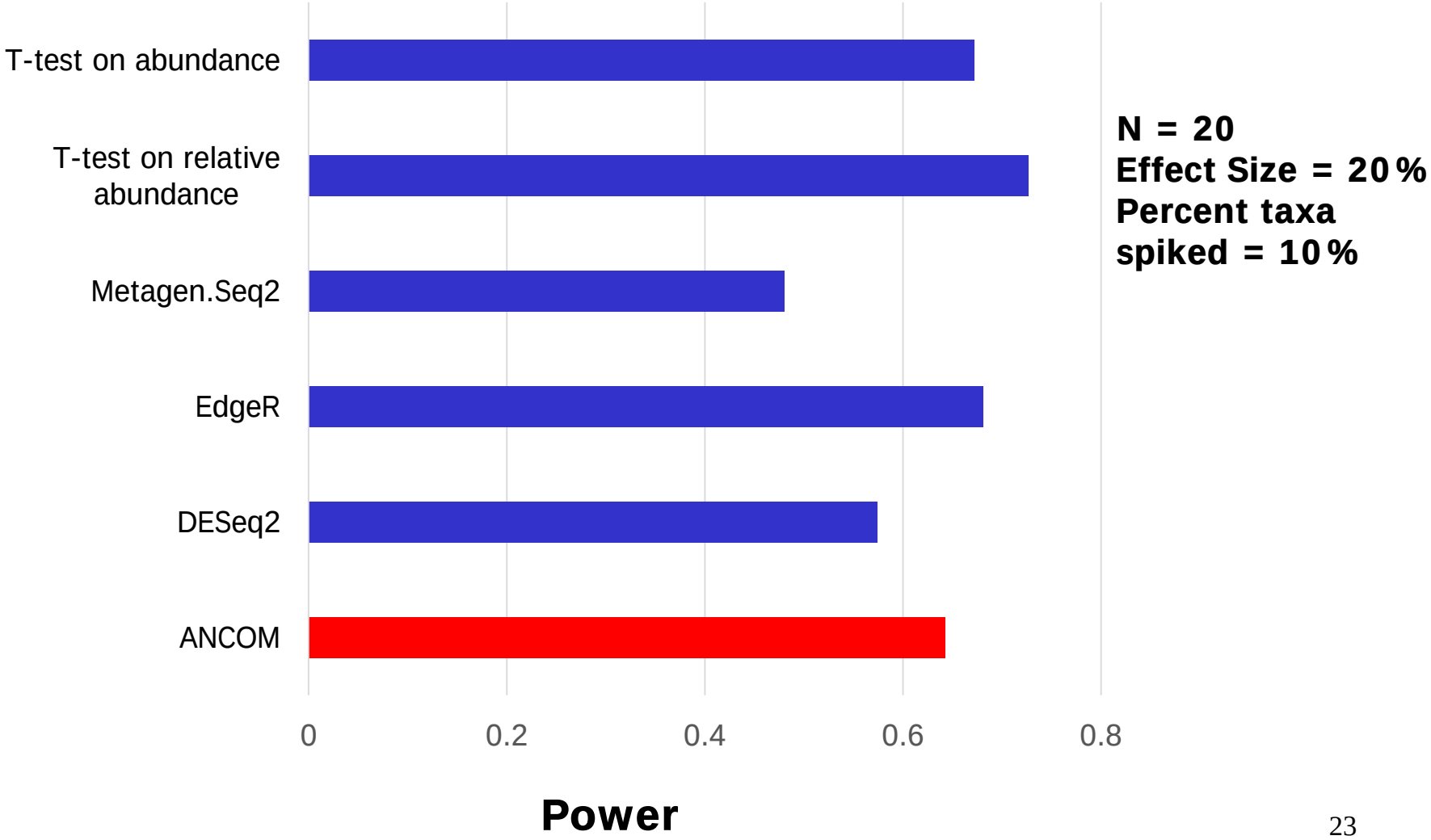
- **Baseline data:** Data on 2000 taxa from the paper
- **Group 1 (control group):** A random sample with replacement is drawn from baseline data
- **Group 2 (treatment group):** A random sample with replacement is drawn from baseline data. For non-null data:
  - Randomly spiked 5, 10, 15 or 20% taxa
  - Amount of spiking 5 to 20% (i.e. increase in abundance)
- **Sample sizes:** 5, 20, 100 per group
- **Number of simulations:** 800
- **FDR nominal level:** 5%.

ANCOM Controls FDR Better Than Other Methods Considered



Weiss et al., 2017

ANCOM Competes Well in Terms of Power



Weiss et al., 2017

Testing problem for more than 2 ecosystems: Constrained inference

Suppose we have 4 ecosystems:

G1: Vaginally born and no antibiotics exposure within the first month

G2: Vaginally born and antibiotics exposure within the first month

G3: C-Section born and no antibiotics exposure within the first month

G4: C-Section born and antibiotics exposure within the first month

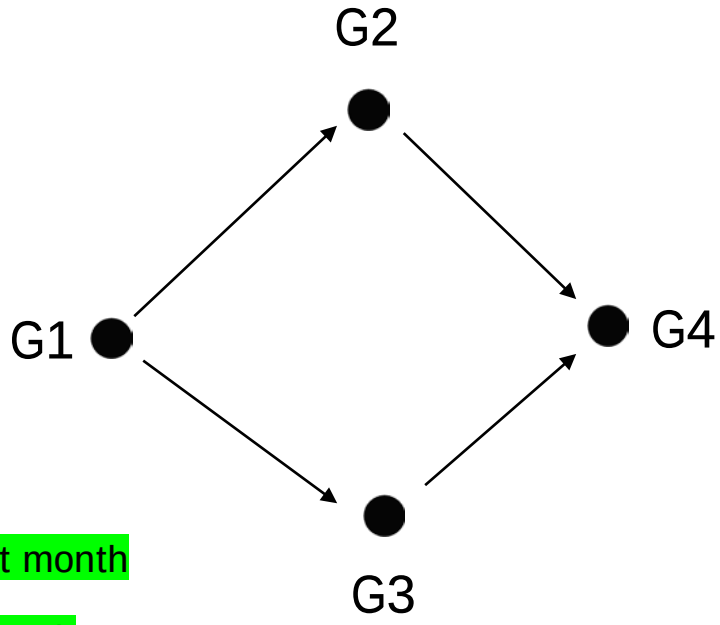
Testing problem for more than 2 ecosystems

Identify taxa whose mean abundance satisfies the following pattern of inequalities for a unit volume of tissue:

$$H_0: E(ln(\mu_j^1)) = E(ln(\mu_j^2)) = E(ln(\mu_j^3)) = E(ln(\mu_j^4))$$

Vs.

$$H_a: \{E(ln(\mu_j^1)) \leq E(ln(\mu_j^2)) \leq E(ln(\mu_j^4))\}$$
$$\cap$$
$$\{E(ln(\mu_j^1)) \leq E(ln(\mu_j^3)) \leq E(ln(\mu_j^4))\}$$



G1: Vaginally born and no antibiotics exposure within the first month

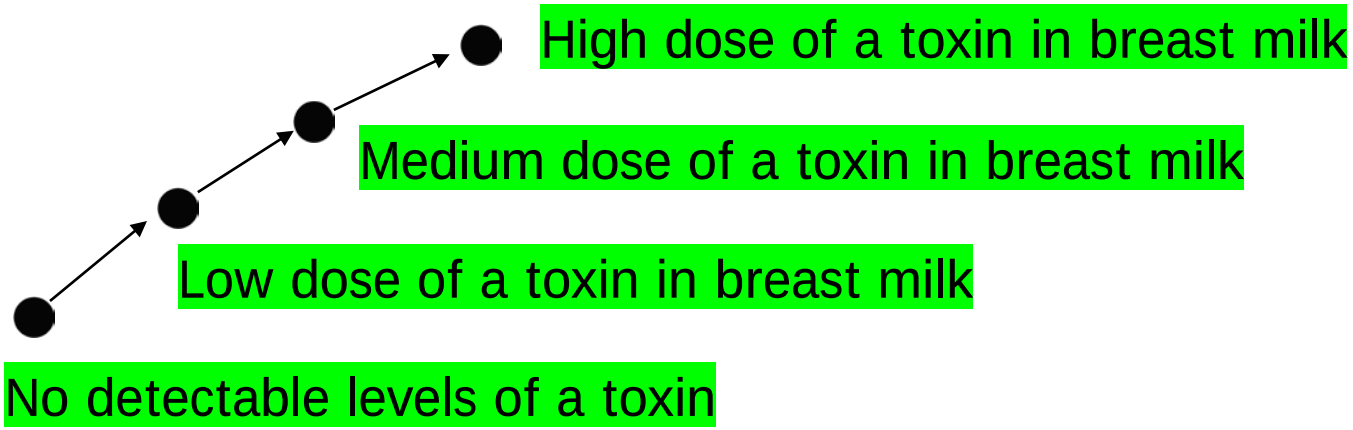
G2: Vaginally born and antibiotics exposure within the first month

G3: C-Section born and no antibiotics exposure within the first month

G4: C-Section born and antibiotics exposure within the first month

Tests for patterns

Identify taxa whose the mean abundance satisfy the following pattern of inequalities for a unit volume of tissue:



$$H_0: E(ln(\mu_j^1)) = E(ln(\mu_j^2)) = E(ln(\mu_j^3)) = E(ln(\mu_j^4))$$

Vs.

$$H_a: \{E(ln(\mu_j^1)) \leq E(ln(\mu_j^2)) \leq E(ln(\mu_j^3)) \leq E(ln(\mu_j^4))\}$$

Constrained inference

Suppose we have 4 ecosystems:

$$H_0: E(ln(\mu_j^1)) = E(ln(\mu_j^2)) = E(ln(\mu_j^3)) = E(ln(\mu_j^4))$$

Vs.

$$H_a: \left(E\left(ln(\mu_j^1)\right), E\left(ln(\mu_j^2)\right), E\left(ln(\mu_j^3)\right), E\left(ln(\mu_j^4)\right)\right)' \in C$$

C: Convex Cone

Using constrained inference methods extend ANCOM to the above problem.

Concluding remarks

- The ANCOM methodology:
  - Provides a better control of false discovery rate than other available methods.
  - For each taxon, it can be extended for testing for patterns among different ecosystems by appealing to constrained likelihood ratio type tests.
  - Can be generalized to covariate adjusted analysis, repeated measurement analysis
  - Software:
    - R code: contact me at sdp47@pitt.edu
    - Python: Available from QIIME2
- Improved version of ANCOM: Visit my student Mr. Huang Lin's poster

Major Collaborators

Siddhartha Mandal (2012 - 2015)
Norwegian Inst. Public Health
Currently: Public Health Foundation of India

Rob Knight, Professor
School of Medicine
Department of Computer Science
and Engineering
UC San Diego, Ca

Merete Eggesbo, Professor
PI: NoMIC Study
Norwegian Institute of Public Health
Oslo, Norway
PI of the NOMIC Study

How does maternal nutrition affect Gut microbiota at delivery?

- Intakes of 28 different nutrients estimated based on Food Frequency Questionnaire in 2nd trimester
- Maternal gut microbiota 4 days after delivery (Illumina)
- **Vitamin D, and to some extent retinol and cholesterol, significantly reduced diversity**
- **Different types of fat (saturated versus monosaturated) could shift composition in opposite direction**

Mandal *et al. Microbiome* (2016) 4:55
DOI 10.1186/s40168-016-0200-3

Microbiome

RESEARCH Open Access

Fat and vitamin intakes during pregnancy have stronger relations with a pro-inflammatory maternal microbiota than does carbohydrate intake

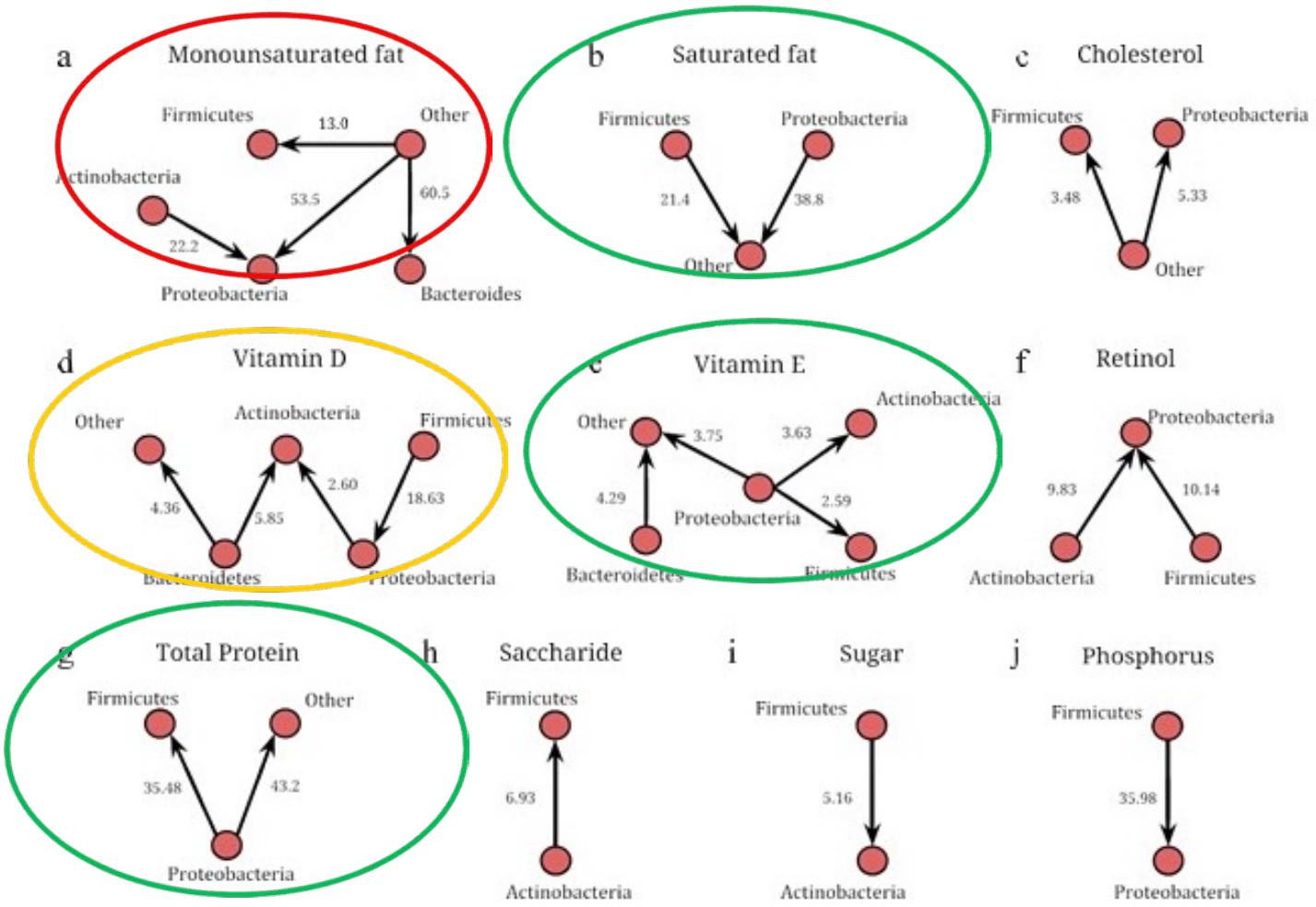
Siddhartha Mandal[1,2], Keith M. Godfrey[3], Daniel McDonald[4], Will V. Treuren[5], Jørgen V. Bjørnholt[1,6,7], Tore Midtvedt[8], Birgitte Moen[9], Knut Rudi[10], Rob Knight[4,11], Anne Lise Brantsæter[1], Shyamal D. Peddada[12] and Merete Eggesbø[1*]

Nat Genet. 2016 Nov;48(11):1396-1406. doi: 10.1038/ng.3695. Epub 2016 Oct 10.

Genome-wide association analysis identifies variation in vitamin D receptor and other host factors influencing the gut microbiota.

Wang J[1,2], Thingholm LB[3], Skiecevičienė J[3], Rausch P[1,2], Kummen M[4,5,6,7], Hov JR[4,5,6,7,8], Degenhardt F[3], Heinsen FA[3], Rühlemann MC[3], Szymczak S[3],

Association between dietary variables and various phyla



Numbers represent the “fold increase” in the taxon (pointed to) relative to the taxon (pointed from) for a unit SD increase in the dietary variable. N= 60 women.

Mandal et al. 2017

More than 2 ecosystems ...

Global test

Suppose there are $G > 2$ ecosystems (or experimental groups) to be compared. A wide range of analyses can be performed

A. Classical global test

$$H_{0,j}: \bigcap_{g_1 \neq g_2}^{G} E(\ln(\mu_j^{g_1}) - \ln(\mu_r^{g_1})) = E(\ln(\mu_j^{g_2}) - \ln(\mu_r^{g_2}))$$

$$H_{a,j}: \bigcup_{g_1 \neq g_2}^{G} E(\ln(\mu_j^{g_1}) - \ln(\mu_r^{g_1})) \neq E(\ln(\mu_j^{g_2}) - \ln(\mu_r^{g_2}))$$

Not a very useful test because rejection of the null only implies there exists at least one ecosystem that is significantly different

Directional tests

B. Directional tests: Often researchers are interested in knowing if the (relative) abundance increased or decreased between two ecosystems for all pairs of ecosystems

$$H_{0,j,g_1,g_2} : E(\ln(\mu_j^{g_1}) - \ln(\mu_r^{g_1})) = E(\ln(\mu_j^{g_2}) - \ln(\mu_r^{g_2}))$$

$$H_{a,j,g_1,g_2} : \left\{E(\ln(\mu_j^{g_1}) - \ln(\mu_r^{g_1})) < E(\ln(\mu_j^{g_2}) - \ln(\mu_r^{g_2}))\right\}$$
$$\cup \left\{E(\ln(\mu_j^{g_1}) - \ln(\mu_r^{g_1})) > E(\ln(\mu_j^{g_2}) - \ln(\mu_r^{g_2}))\right\}$$

Total number of hypotheses to be tested = $2^{\binom{G}{2}} \times (m-1)\binom{G}{2}$

Directional tests

B. Directional tests:

- BH procedure for the above multiple testing problem will be too conservative

- Instead one can use mdFDR controlling procedure of Guo et al. (2010). It controls the overall FDR (under the same assumptions as BH procedure while being substantially more powerful than BH

Step1: For each taxon, perform the following two-sided test, using t-test

$$H_{0,j,g_1,g_2} : E(\ln(\mu_j^{g_1}) - \ln(\mu_r^{g_1})) = E(\ln(\mu_j^{g_2}) - \ln(\mu_r^{g_2}))$$

$$H_{a,j,g_1,g_2} : E(\ln(\mu_j^{g_1}) - \ln(\mu_r^{g_1})) \neq E(\ln(\mu_j^{g_2}) - \ln(\mu_r^{g_2}))$$

Let p_{j,g_1,g_2} denote the corresponding p-value

Directional tests

Step2: Let $\tilde{p}_j = \binom{G}{2} \min_{g_1,g_2} \{p_{j,g_1,g_2}\}$

Step 3: Apply BH procedure on the adjusted p-values $\tilde{p}_j$, $j = 1,2,...,m$ at a pre-specified level of significant α

Step 4: Suppose R null hypotheses are rejected out of total m hypotheses in Step 3

Step 5: For every taxon j declared significant in Step 4 with $p_{j,g_1,g_2} \leq \frac{R}{m\binom{G}{2}}\alpha$, if $T_{j,g_1,g_2} > (<)0$ then declare that

$$E(\ln(\mu_j^{g_1}) - \ln(\mu_r^{g_1})) > (<) E(\ln(\mu_j^{g_2}) - \ln(\mu_r^{g_2}))$$

Tests against a specific ecosystem

C. Directional tests against a prespecified ecosystem (e.g. Control group): Often researchers are interested in knowing if the (relative) abundance increased or decreased in an ecosystem relative a pre-specified ecosystem.

$$H_{0,j,g,control} : E(\ln(\mu_j^g) - \ln(\mu_r^g)) = E(\ln(\mu_j^{control}) - \ln(\mu_r^{control}))$$

$$H_{a,j,g,control} : \left\{E(\ln(\mu_j^g) - \ln(\mu_r^g)) < E(\ln(\mu_j^{control}) - \ln(\mu_r^{control}))\right\}$$
$$\cup \left\{E(\ln(\mu_j^g) - \ln(\mu_r^g)) > E(\ln(\mu_j^{control}) - \ln(\mu_r^{control}))\right\}$$

Total number of hypotheses to be tested = $2^{(G-1)} \times (m-1)(G-1)$

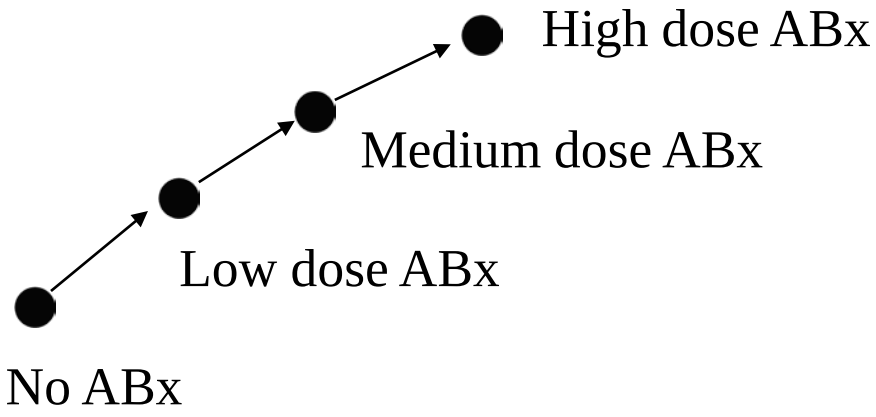
Tests against a specific ecosystem

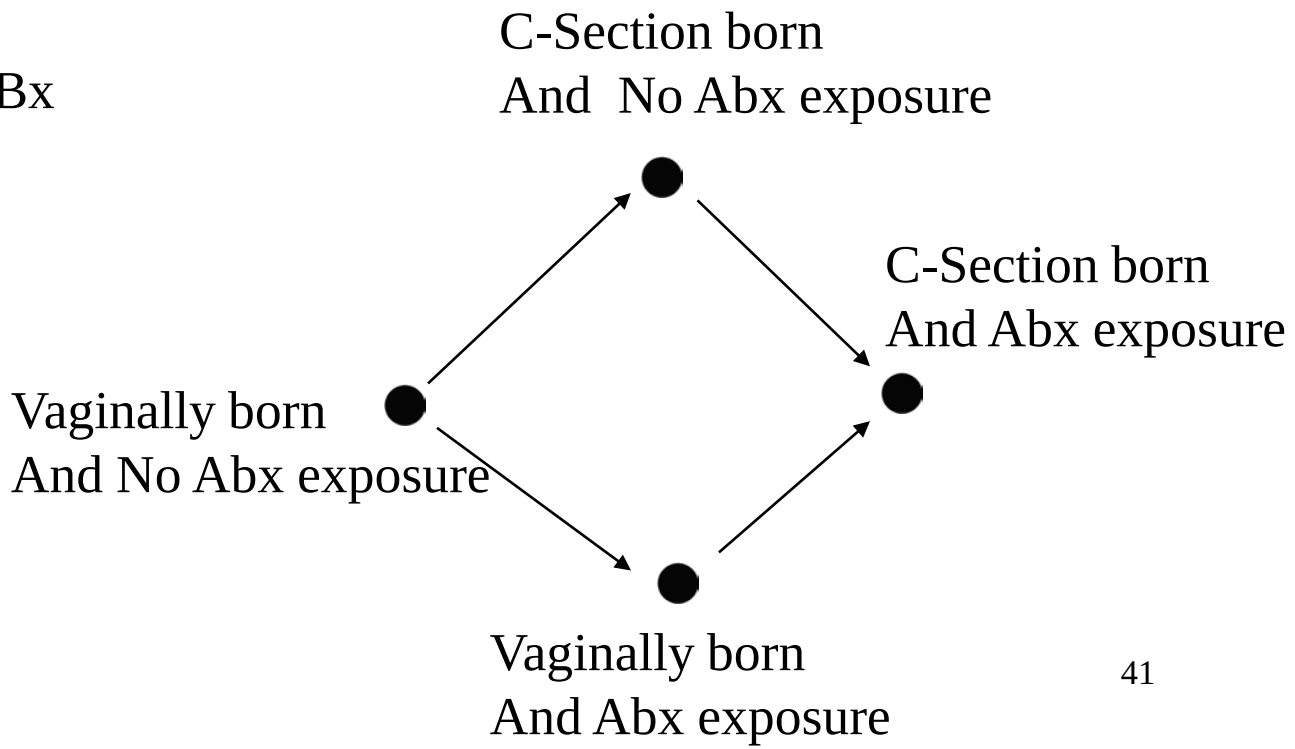
C. Directional tests:

Instead of mdFDR controlling procedure of Guo et al. (2010) one can use a generalization of Dunnett's type test of Grandhi et al. (2016) which is more powerful than Guo et al. (2010)

Tests for patterns

D. Test for trends or patterns:

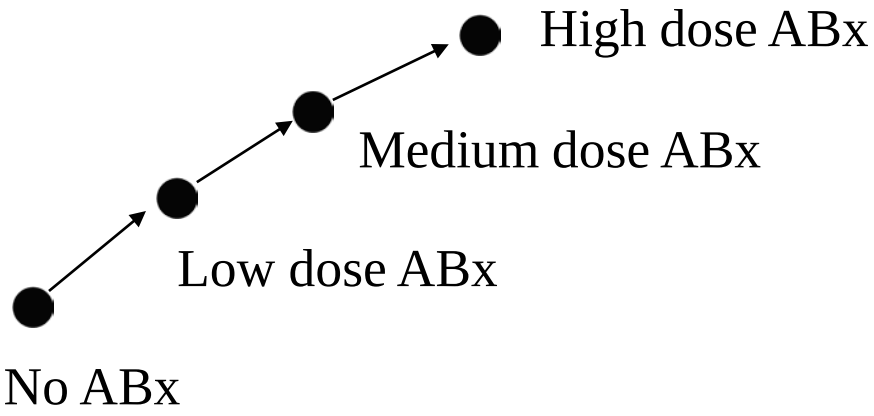




Kaul et al. (2017)

Tests for patterns

D. Test for trends or patterns:

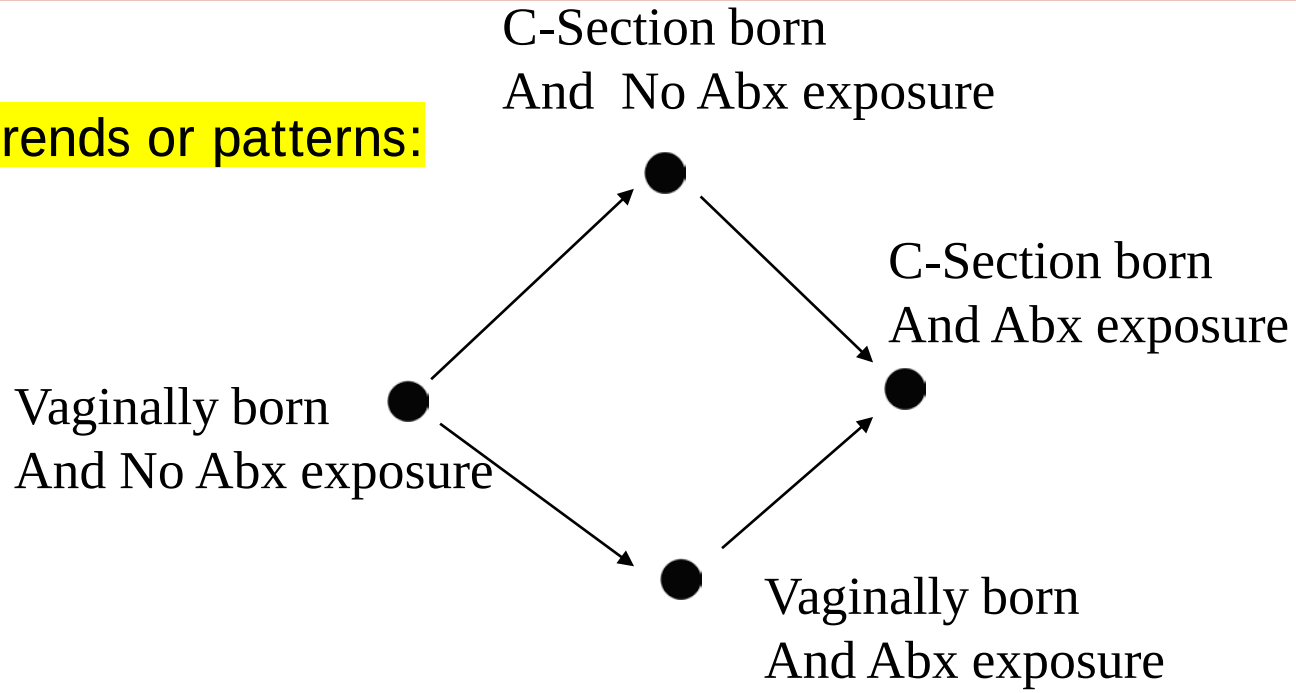


$$H_{0,j} : E(\ln(\mu_j^1) - \ln(\mu_r^1)) = E(\ln(\mu_j^2) - \ln(\mu_r^2)) = \ldots = E(\ln(\mu_j^G) - \ln(\mu_r^G))$$

$$H_{a,j} : E(\ln(\mu_j^1) - \ln(\mu_r^1)) \leq E(\ln(\mu_j^2) - \ln(\mu_r^2)) \leq \ldots \leq E(\ln(\mu_j^G) - \ln(\mu_r^G))$$

Tests for patterns

D. Test for trends or patterns:



$$H_{0,j} : E(\ln(\mu_j^1) - \ln(\mu_r^1)) = E(\ln(\mu_j^2) - \ln(\mu_r^2)) = E(\ln(\mu_j^3) - \ln(\mu_r^3))$$
$$= E(\ln(\mu_j^4) - \ln(\mu_r^4))$$

$$H_{a,j} : \left\{E(\ln(\mu_j^1) - \ln(\mu_r^1)) \le E(\ln(\mu_j^2) - \ln(\mu_r^2)) \le E(\ln(\mu_j^4) - \ln(\mu_r^4))\right\}$$
$$\cup \left\{E(\ln(\mu_j^1) - \ln(\mu_r^1)) \le E(\ln(\mu_j^3) - \ln(\mu_r^3)) \le E(\ln(\mu_j^4) - \ln(\mu_r^4))\right\}$$

Tests for patterns

D. Test for trends or patterns:

More generally one can test union of all patterns of interest using the order restricted inference based methods of Peddada et al. (2003), Farnan et al. (2014), Jelsema and Peddada (2016)

$$H_{0,j} : E(\ln(\mu_j^1) - \ln(\mu_r^1)) = E(\ln(\mu_j^2) - \ln(\mu_r^2)) = \ldots = E(\ln(\mu_j^G) - \ln(\mu_r^G))$$

$$H_{a,j} : \text{Union of all patterns of interest}$$

Kaul et al. (2017)