A Powerful Two-stage Microbiome-wide Association test

Huilin Li Ph.D.
Division of Biostatistics
Department of Population Health

BIRS 2019 Feb. 4-8

Genomics and Metagenomics workshop

Human Microbiome

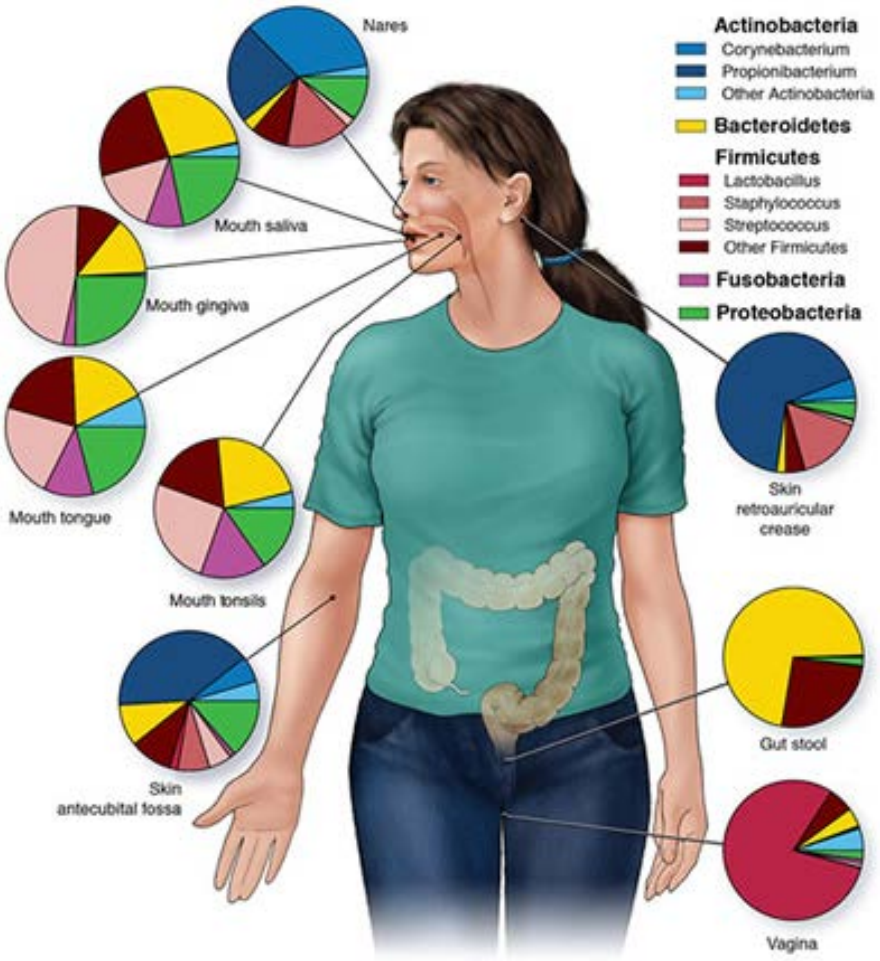
- The communities of microbes living in and on the various parts of your body

- Function of microbial community
  - Digestive enzyme
  - Metabolism of food constituents
  - Protection from pathogens.
  - Interaction with the immune system

Picture source: Synbiocyc.org

Human Microbiome

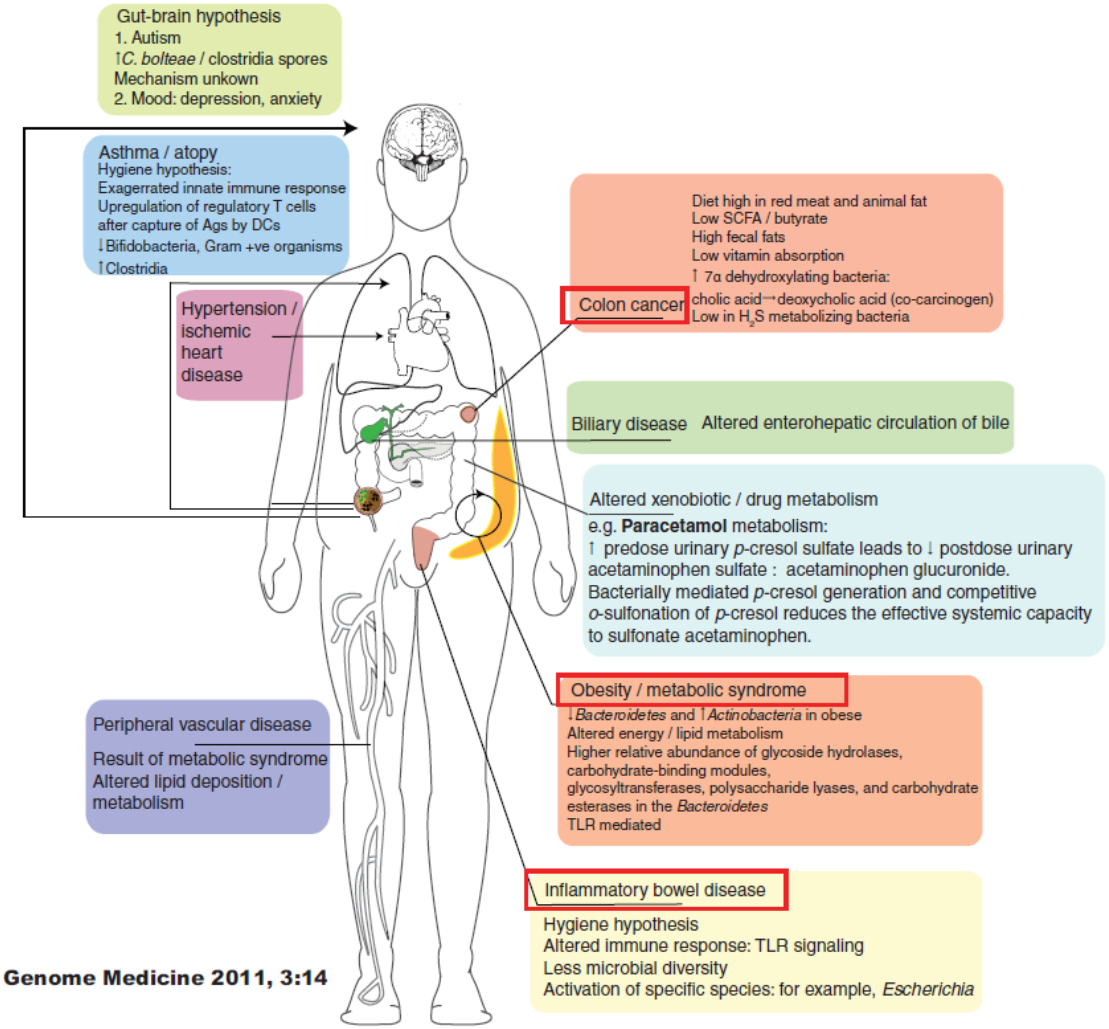
- Bigger variation than the human genome
- Personal; Distinctive microbial profile at different body sites
- Microbial state often differs in health and disease
- Restore the “out of balance” microbial profile to normal



Cho and Blaser 2012

Picture source: allergiesandyourgut.com

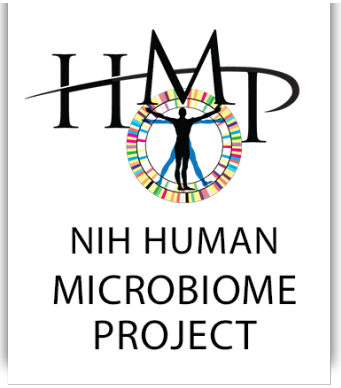
Microbiome and Human Disease



Kinross et al. Genome Medicine 2011, 3:14

Experimental Design

- **Cross Sectional Studies**
  - **Finding differences in microbial communities between different human populations**
- **Randomization Trial**
  - **Identifying the treatment effect**
- **Longitudinal Studies**
  - **Investigating the stability and dynamics of microbial communities**



Statistical Analysis

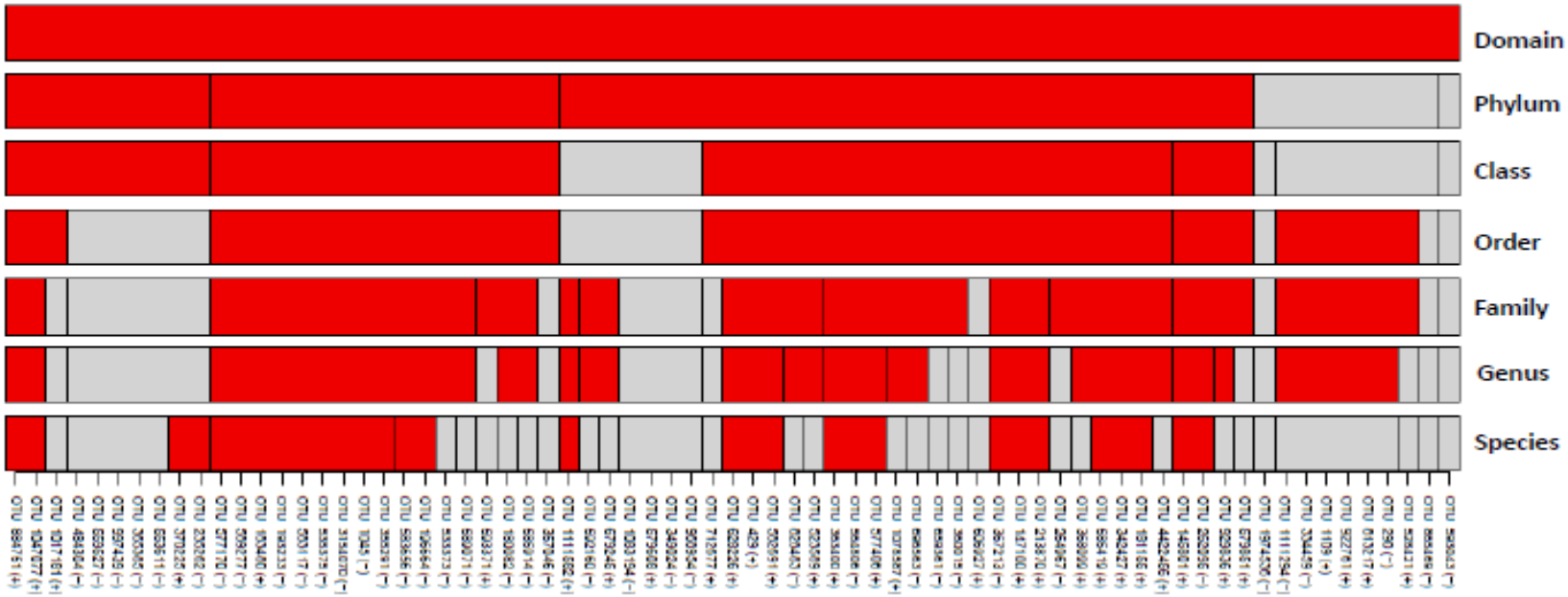
- Community level analyses
- Taxonomical level analyses
- Advanced analysis in longitudinal study
  - Microbial dynamic modeling
  - Survival analysis(time-to-event outcome)
  - Causal/Mediation analysis

Microbiome-wide Association Study(MWAS)

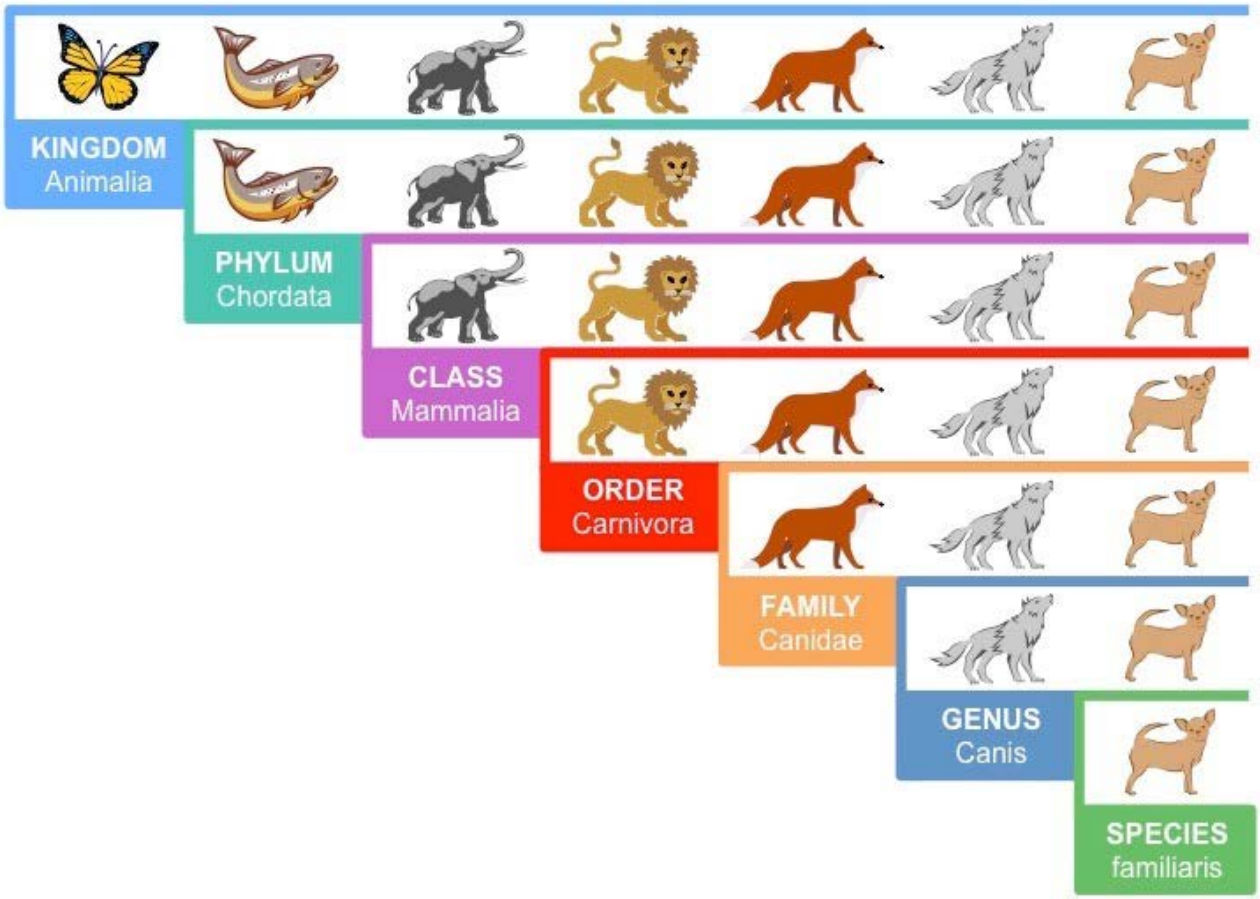
- In microbiome studies, MWAS is a study of a microbiome-wide set of taxa live in different individuals to see if any taxa is associated with a trait.
- Trait could be:
  - Binary outcome — disease status
  - Continuous outcome-- clinical biomarker(e.g. CD4+, BMI,…)
  - Survival outcome—time to T1D onset, time to recurrence etc.

Structure of Microbiome Data

- All strains in the **domain *Bacteria*** in mammal are hierarchically classified into six major levels
- At each lower level, organisms are classified with their most similar cousins based on common features



Taxonomic Classification



Ib. bioninja.com.au

Microbiome Data

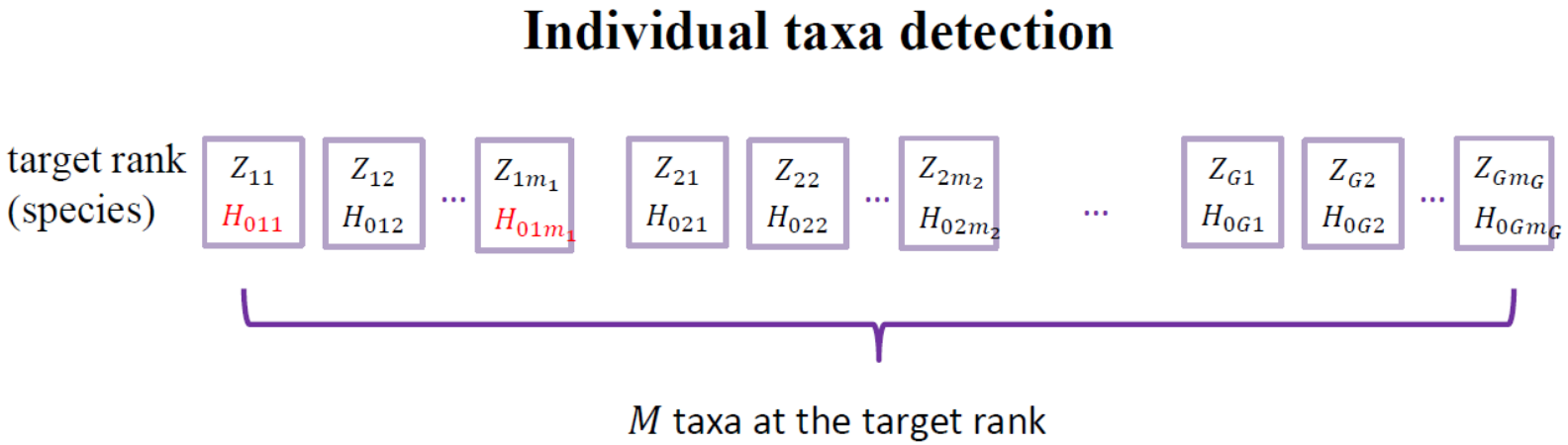
There are three components:

1. Relative abundance table: $Z_{n \times p}$
2. Tree information:
   - taxonomic tree: group classification
   - phylogenetic tree distance matrix $\mathrm{D}_{p \times p}$
3. Other covariates, trait or outcome $\mathrm{X}_{n \times m}$

Traditional one-stage method

Test the association for microbes individually and utilize BH procedure afterwards to control the FDR

- Problems:
  - Assume independency of hypotheses
  - Large number of multiple comparison– very few discovery

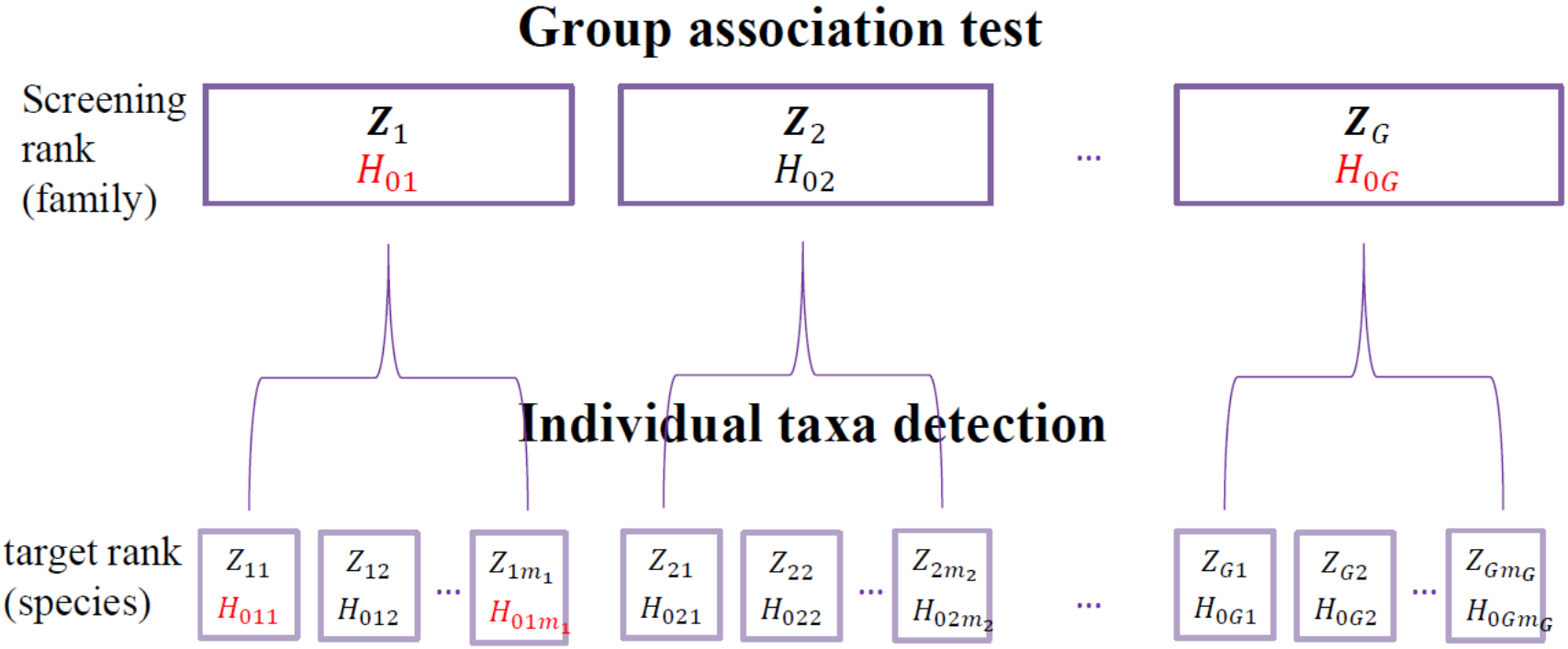


Motivation for a Two-stage Test

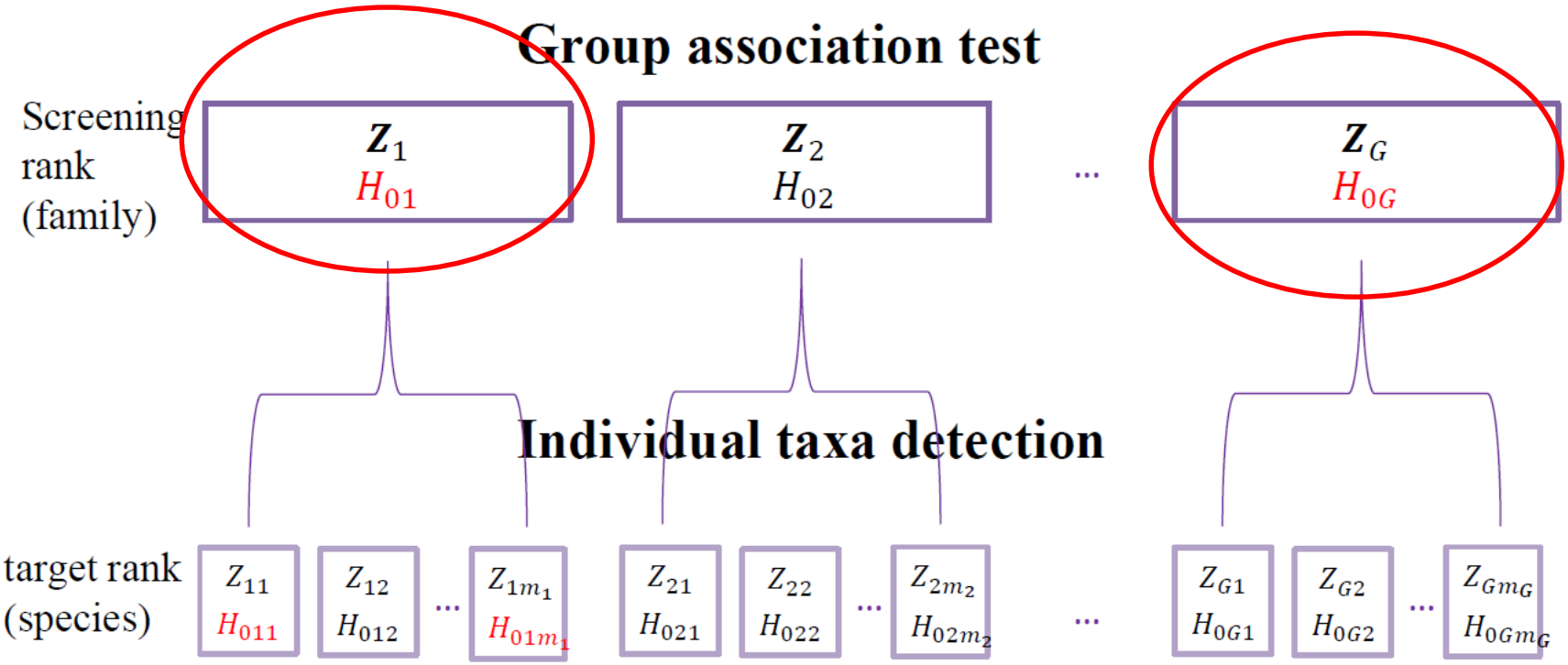
- The trait-associated taxa tend to be clustered evolutionarily instead of randomly distributed across the community
- The known taxonomic structure depicts the microbial evolutionary relationships

A new test which incorporates the prior biological information through the **taxonomic tree** to alleviate multiplicity issue, thus enhance the statistical power

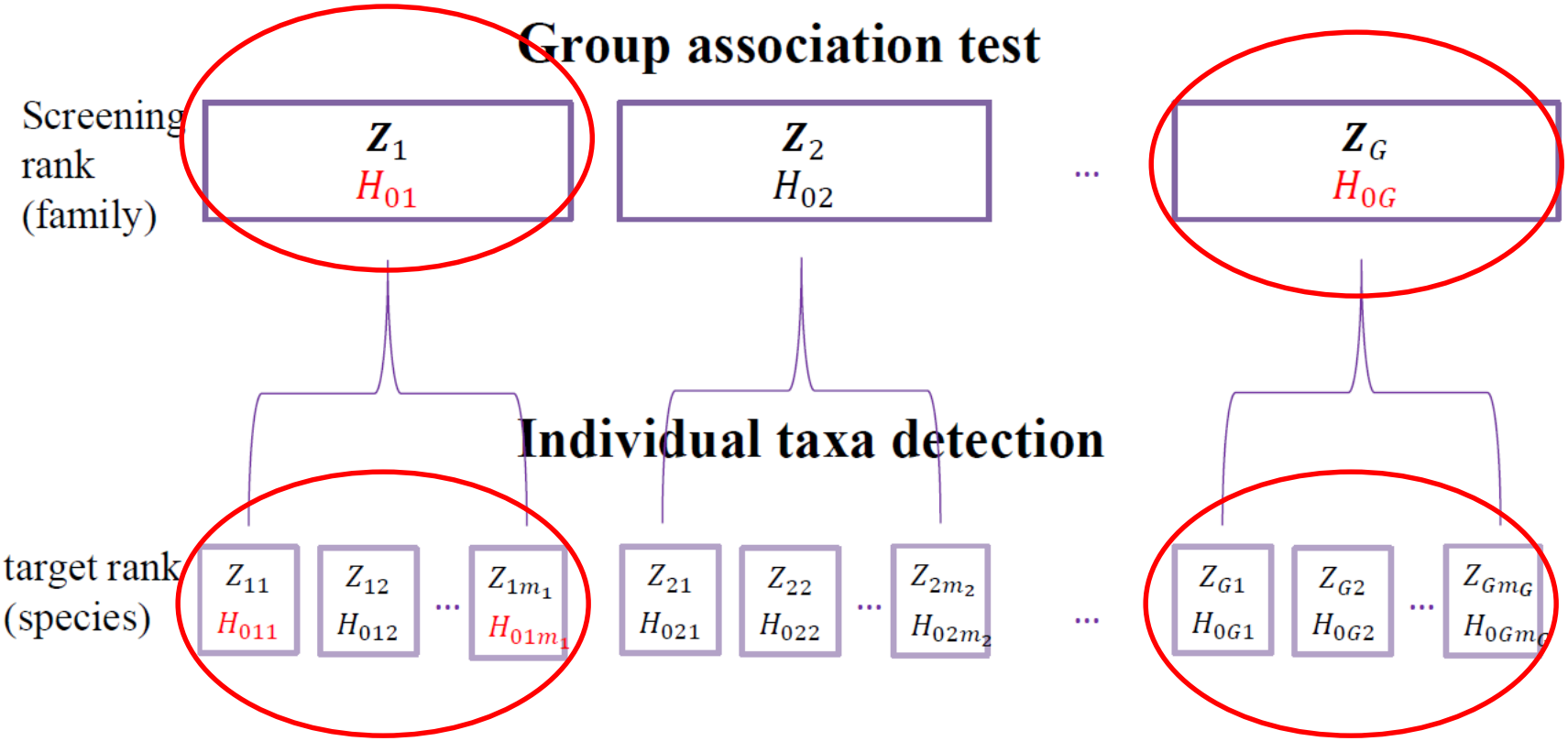
A Two-stage Microbial Association Mapping Framework (massMap)



A Two-stage Microbial Association Mapping Framework (massMap)



A Two-stage Microbial Association Mapping Framework (massMap)



Three Building Components for massMap

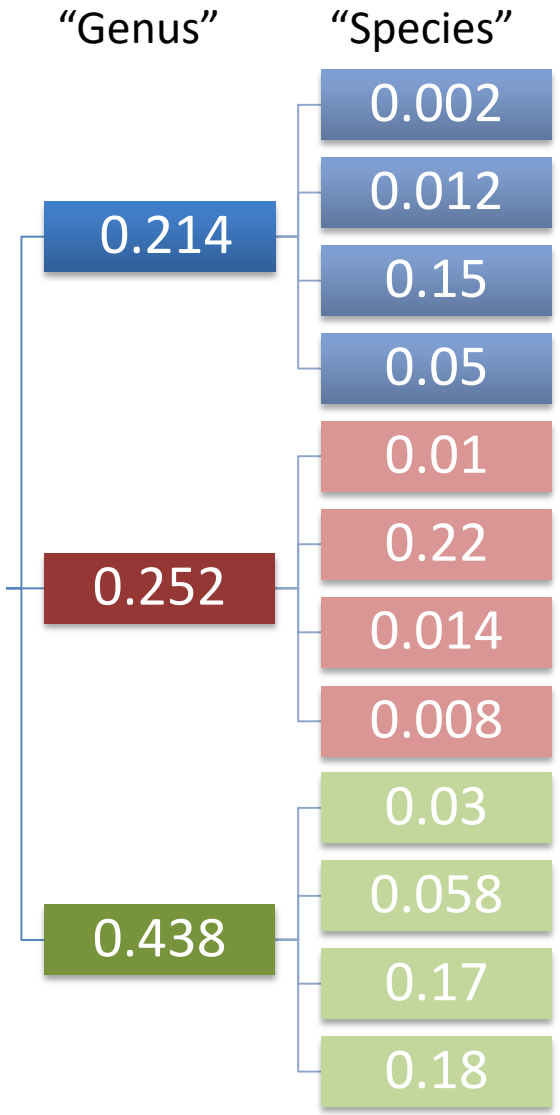
- A powerful **microbial group test** to identify the taxonomic groups that contain the associated taxa
  - OMiAT—Binary and continuous outcomes
  - OMiSA—Survival outcome
- A **pre-selected taxonomic rank** for screening
- An **advanced FDR-controlling** methodology to resolve the dependency among taxa

The Conventional Microbial Association Test

Two Steps:

1. Calculate the relative abundances for the upper level taxa as the aggregates in the lower level lineage
2. Test the association for microbes one by one at each rank and utilize BH procedure afterwards to control the FDR

We call those methods as the **aggregate-based methods**.



The aggregate-based method

- **Assumption:** the associated microorganisms nested in each upper-level taxon are **all in the same effect direction.**
- **Problem:** This approach is inefficient by neglecting detailed information about diverse association patterns from nested microorganisms

Examples:

- ✓ LEfSe (Segata et al. 2011)
- ✓ metagenomeSeq-fit Zig (Paulson et al. 2013)
- ✓ STAMP (Parks et al, 2014)

Microbial Group Association Test

- Y_i and $\boldsymbol{X}_i$ denote the binary outcome trait and covariates for subject i
- $\boldsymbol{Z}_{ig} = \left(Z_{ig1}, Z_{ig2}, \ldots, Z_{igm_g}\right)'$ is the relative abundance of taxa in the gth group

- $\text{Logit}[\text{P}(Y_i = 1)] = \beta_0 + \boldsymbol{\alpha}'\boldsymbol{X}_i + \boldsymbol{\beta}_g'\boldsymbol{Z}_{ig}$
- $\boldsymbol{\beta}_g = \left(\beta_{g1}, \beta_{g2}, \ldots, \beta_{gm_g}\right)'$ is the vector of coefficients for taxa from group g

$$H_{0g}: \beta_{g1} = \beta_{g2} = \cdots = \beta_{gm_g} = 0$$
$$v.s.\ H_{1g}: \text{at least one } \beta_{gj} \neq 0, \quad j = 1, \ldots, m_g$$

The diverse association patterns

- ✓The associated taxa have **the same effect direction**.
- ✓The associated taxa are in **mixed effect direction**.
- ✓The **abundant** taxa are associated.
- ✓The **rare** taxa are associated.
- ✓The **phylogenetic tree distance**

Omnibus Microbiome Association Test (OMiAT)

- OMiAT: $M^g_{OMiAT} = \text{minP}\{T^g_{aSPU}, Q^g_{OMiRKAT}\}$.

  - T^g_{aSPU} is useful for modulating different association patterns arising from highly imbalanced microbial abundances. (Pan et al. 2014)

  - $Q^g_{OMiRKAT}$ is advantageous in detecting microbial group associations utilizing phylogenetic tree information, is tailored from the microbiome regression-based kernel association test (MiRKAT)[27],

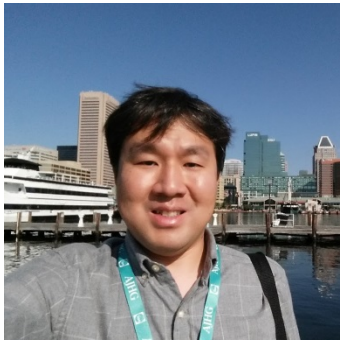
- **Features:**
  - **A data-driven approach.**
  - **Highly robust and powerful.**

Omnibus Microbiome Association Test (OMiAT)

- OMiAT: Koh, H. et al. Microbiome. 2017;5:45

  - ❖ It is a powerful test specifically designed for the detection of varying association patterns at the higher taxonomic rank
  - ❖ It can accommodate multiple covariates
  - ❖ It is a useful screening test

- **Software: OMiAT**
  - **https://sites.google.com/site/huilinli09/software**



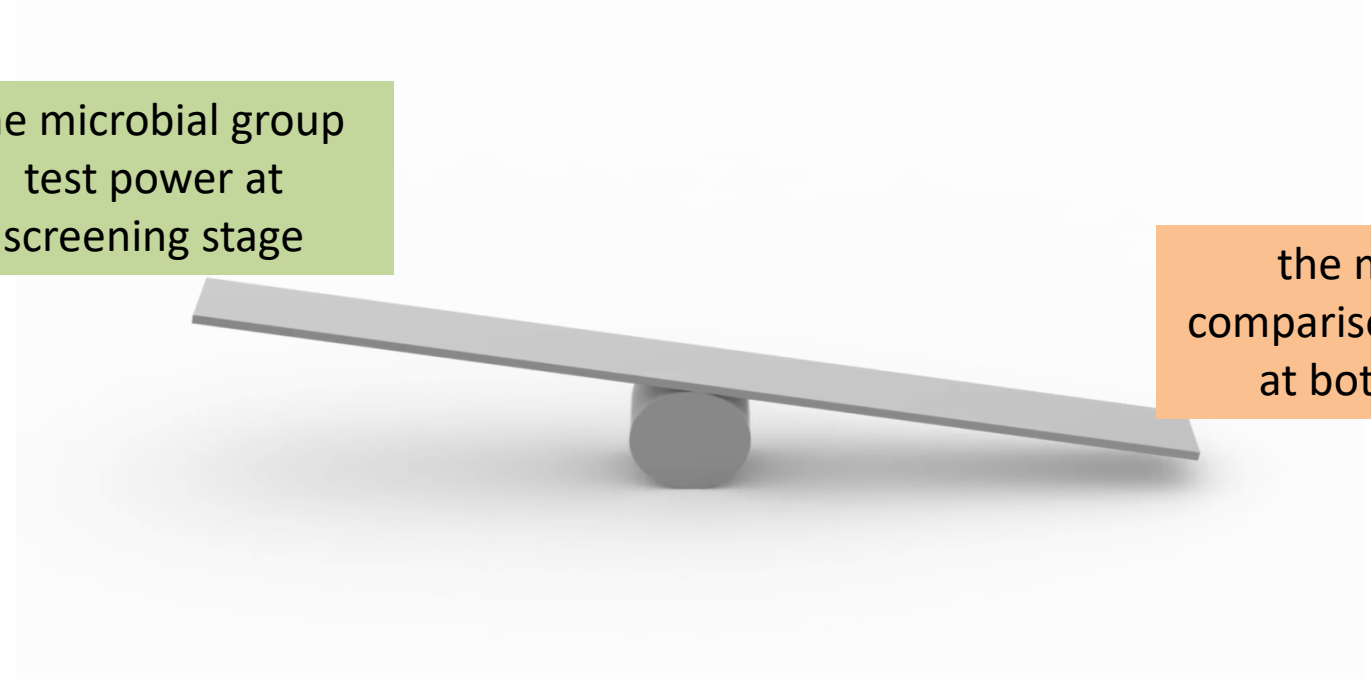
Dr. Hyunwook Koh

Omnibus Microbiome-based Survival Analysis (OMiSA)

- ✓ **Optimal Microbiome-based Survival Analysis (OMiSA), which includes**
  - Optimal Microbiome-based Survival Analysis using Linear and Non-linear bases of OTUs (OMiSALN),
  - Optimal Microbiome Regression-based Kernel Association Test for Survival traits (OMiRKAT-S).
- ✓ **Software: OMiSA**
  - https://sites.google.com/site/huilinli09/software
- ✓ **Reference**
  - Koh, H, Livanos, AE, Blaser, MJ, and Li, H.(2018) A highly adaptive microbiome-based survival analysis method. BMC Genomics.

Which rank to perform the screening?

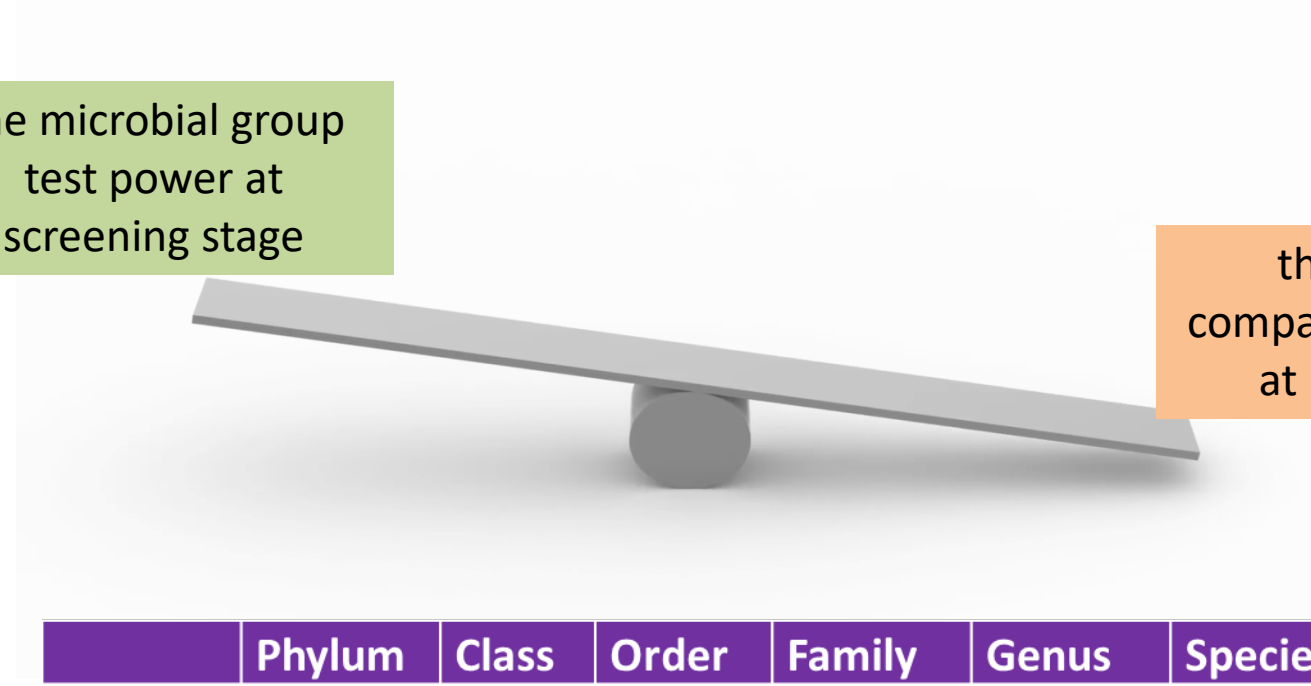
the microbial group test power at screening stage



the multiple comparison penalties at both stages

Which rank to perform the screening?

the microbial group test power at screening stage

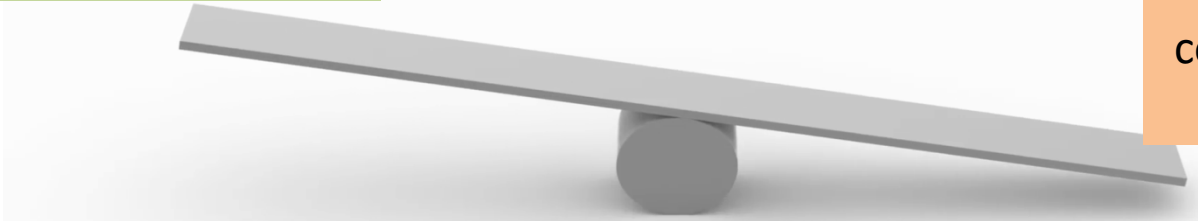


the multiple comparison penalties at both stages

	Phylum	Class	Order	Family	Genus	Species
# groups	6	13	20	41	70	**90**
Group size	15.00	6.92	4.50	2.20	1.29	

Which rank to perform the screening?

the microbial group test power at screening stage



the multiple comparison penalties at both stages

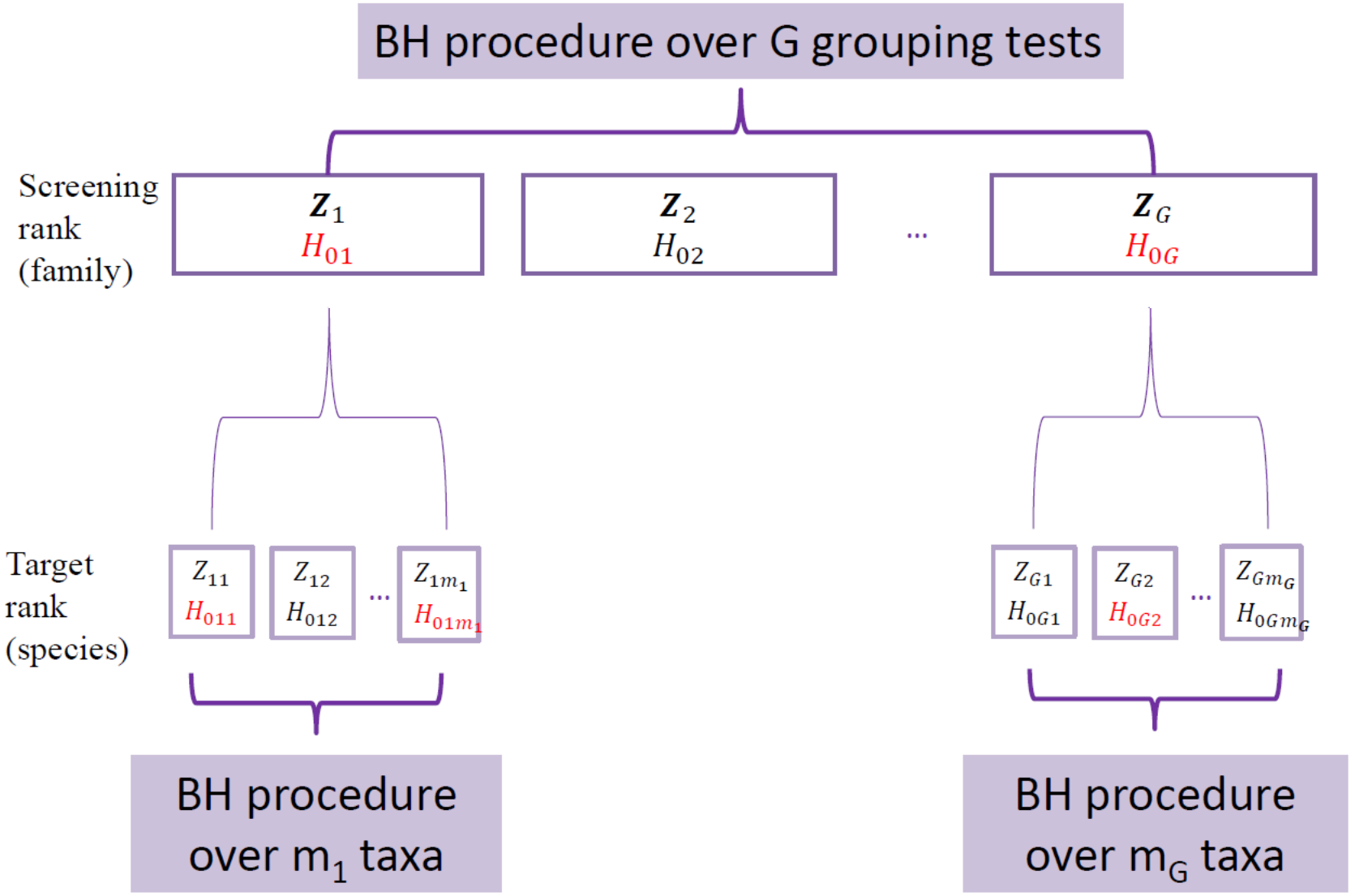
A middle taxonomic rank such as order or family is expected to perform best in the proposed two-stage framework.

Advanced FDR controlling procedures

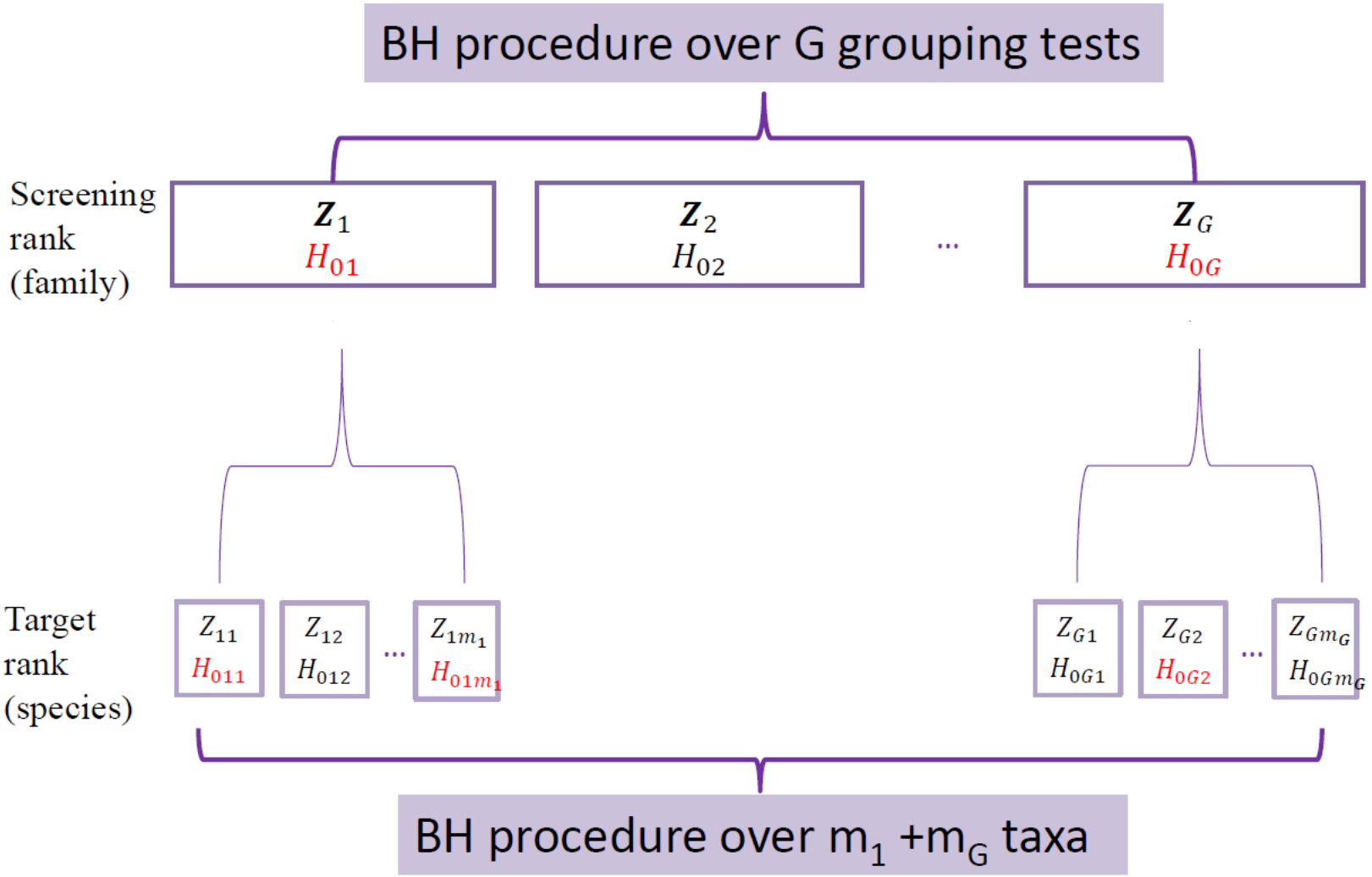
Two advanced FDR-controlling procedures to accommodate the hierarchically structured hypotheses in massMap.

- The hierarchical BH (HBH) procedure (Yekutieli et al. 2006)
- The selected subset testing with BH (SST) procedure (Benjamini and Yekutieli 2005)

The Hierarchical BH (HBH) procedures



The Selected Subset BH (SST) Procedures



Simulations

- Simulated OTU counts for **200 subjects** from the **DM distribution.**
- Total reads =15,000 for sample.
- The dispersion parameter and proportion means. - Estimated from a real microbiome data (AGP data) for **174 OTUs** with original taxonomic tree.
- Generated binary outcome values.

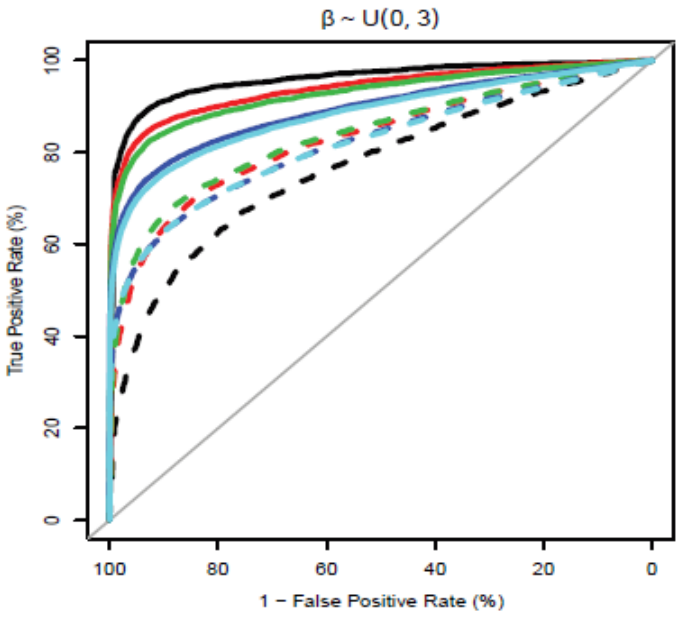
$$\text{Logit}\,[\mathrm{P}(Y_i = 1|\boldsymbol{Z}_i)] = \sum_{j\in\Lambda}\beta_j \text{scale}(Z_{ij})$$

- Partitioned all OTUs into 10 clusters using **PAM** algorithm.

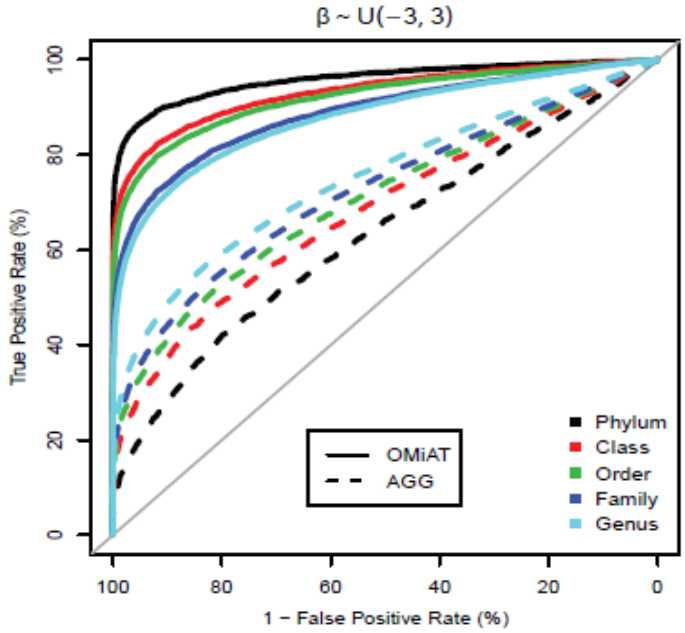
  Randomly set 10% OTUs in 2-3 PAM clusters as the associated OTUs.

Simulation Results

The screening performance of OMiAT and the aggregated method



AUC (%)	Phylum	Class	Order	Family	Genus
OMiAT	95.84	93.78	92.68	89.18	88.39
AGG	76.75	82.88	83.78	81.92	81.78

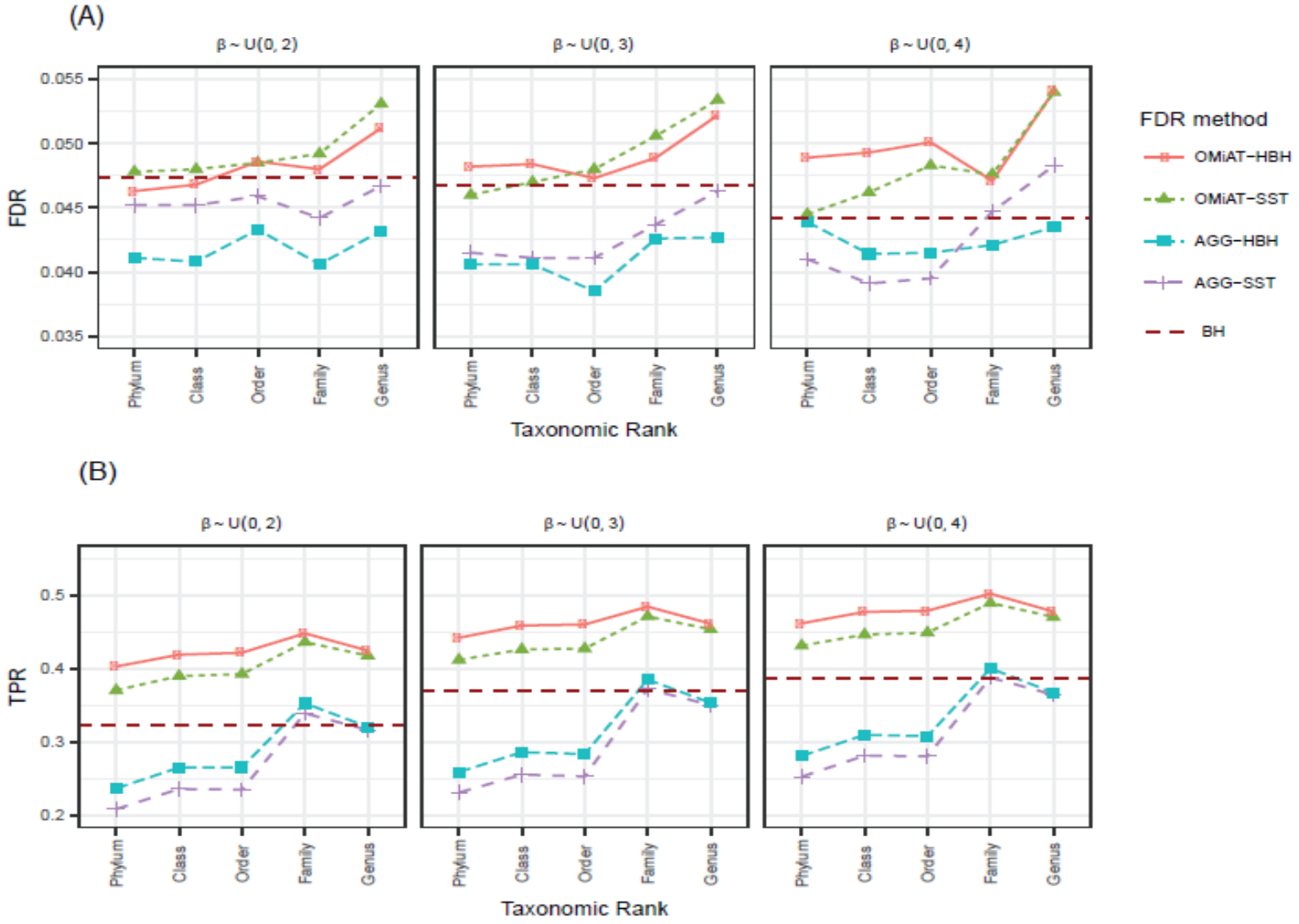


AUC (%)	Phylum	Class	Order	Family	Genus
OMiAT	95.83	93.06	92.01	88.80	87.83
AGG	63.21	68.32	70.64	72.39	74.91

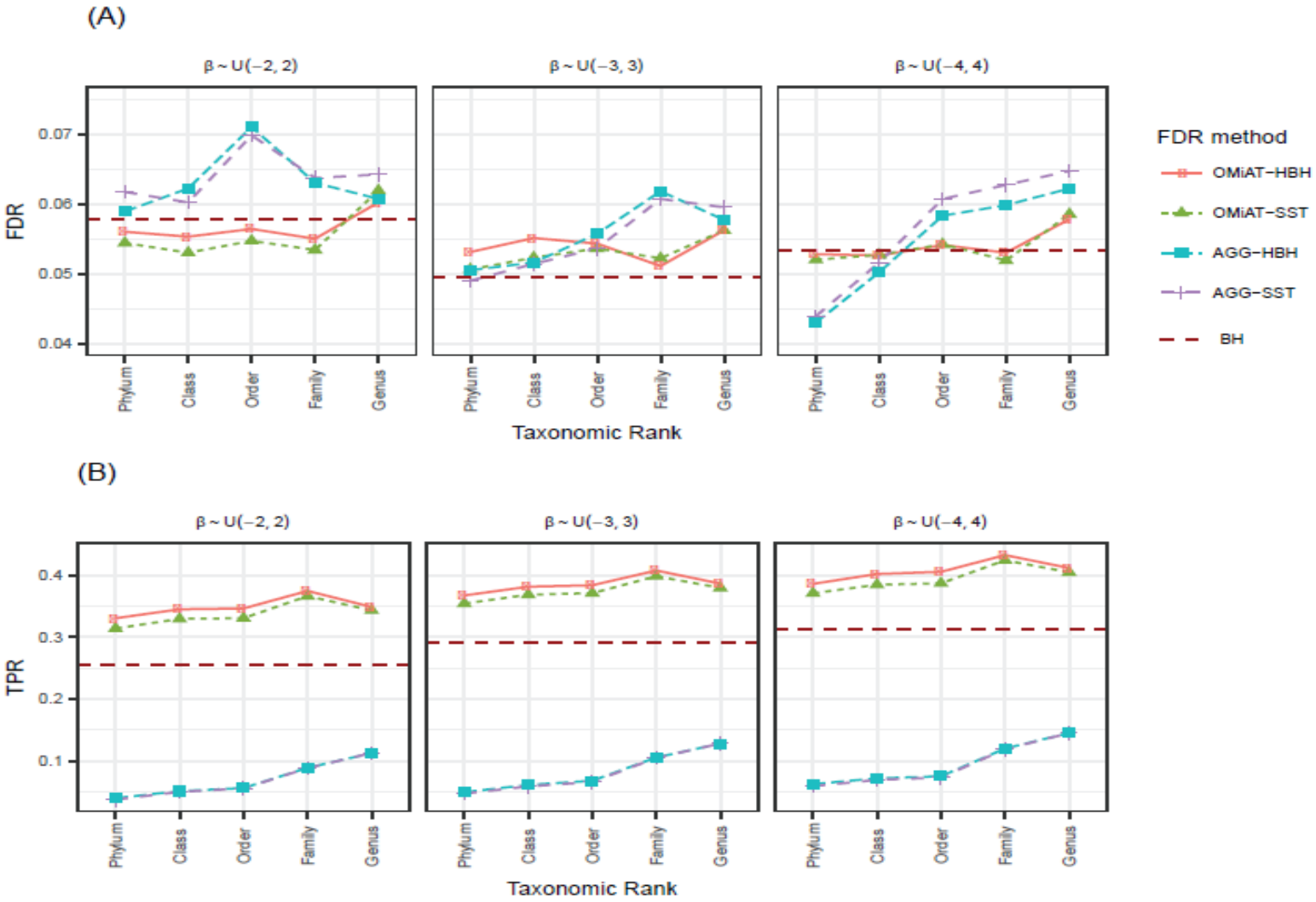
Simulations

- For those 17 associated taxa, we considered two scenarios of association.
  - Under scenario 1, effects of associated taxa have the same sign but varied strength, with small ($\beta_j \sim$ Uniform (0, 2)), modest ($\beta_j \sim$ Uniform (0, 3)) or large effect sizes ($\beta_j \sim$ Uniform (0, 4)).
  - Under scenario 2, the effect directions were mixed in scenario 2, i.e., $\beta_j \sim$ Uniform (-2, 2), Uniform (-3,3), or Uniform(-4, 4).

Results: the Empirical FDR and TPR at the Target Rank(Scenario 1)



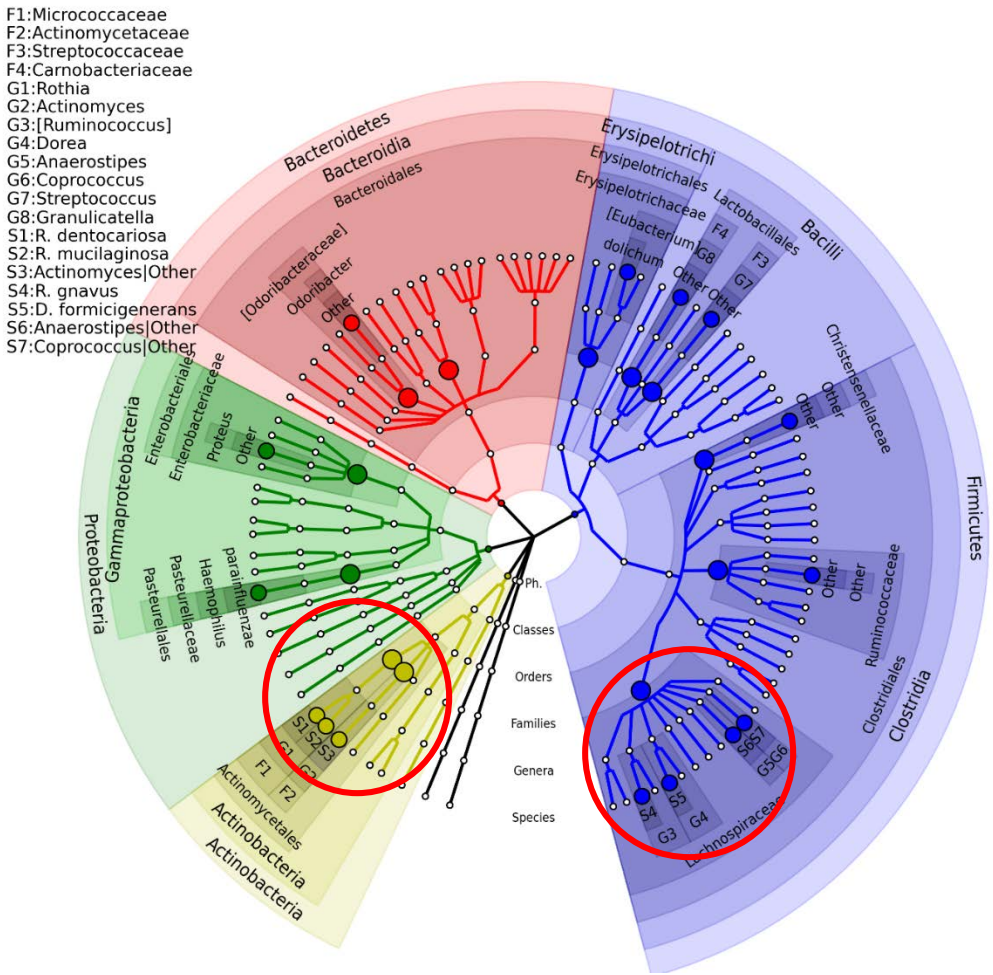
Results: the Empirical FDR and TPR at the Target Rank(Scenario 2)



Real Data Analysis -- American Gut Project

- The American Gut Project aims to create a comprehensive map of the human microbiome.
- 7,293 subjects, 456 descriptive variables, 22,891 OTUs
- After filtering: **1147** samples & **90 species** left for investigation
- Two traits of interest:
  - Antibiotic history (ABH)
  - Body mass index (BMI)
- Covariates: age, gender
- Screening rank: family

AGP—Antibiotic History (ABH)



FDR = 0.05

- Highly overlapping results with competing methods
- Much smaller adjusted p-values
- Clustering association pattern observed – consistent with our assumption

AGP—BMI

OTU ID	Species	Raw p-value	BH	OMiAT-HBH	OMiAT-SST
297635	[Eubacterium] biforme	1.90E-04	1.70E-02	7.60E-04	2.50E-03
824876	Bifidobacterium| Other	2.70E-03		5.30E-03	1.70E-02
4319938	Clostridiaceae| Other	1.00E-02		2.00E-02	3.50E-02
840279	[Barnesiellaceae]|Other	1.10E-02		1.10E-02	3.50E-02
4480861	Catenibacterium| Other	1.50E-02		3.10E-02	4.00E-02
513664	Prevotella stercorea	2.00E-02		8.00E-02	4.30E-02
Number of detected BMI-associated species			1	6	6

Summary

- We develop a two-stage microbial association mapping framework -- **massMap** for binary, continuous and survival outcomes.
- MassMap incorporates the highly powerful microbial group test OMiAT/OMiSA for screening and HBH/SST for the control of FDR.
- A highly efficient method for microbiome-wide association analyses

Dr. Jiyuan Hu

Acknowledgements

Our group members:

Postdoc:

Dr. Jiyuan Hu

Dr. Chan Wang

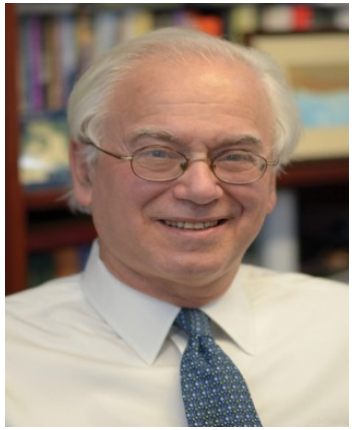
Ph.D. student:

Ms. Linchen He

The research is funded by NIH R01 DK110014.

Acknowledgements

Collaborators:



Dr. Martin Blaser and people in his lab

Dr. Zhiheng Pei and people in his lab

Dr. Yu Chen and people in her lab

Dr. Jiyoung Ahn and people in her lab

…

THANK YOU

NYU Langone Health

References

- Zuur AF, Ieno EN, Walker NJ, Saveliev AA, Smith GM. Mixed Effects Models and Extensions in Ecology with R. New York, NY: Springer Science & Business Media, LLC; 2009.
- Martín-Fernández JA, Hron K, Templ M, Filzmoser P, Palarea-Albaladejo J. Model-based replacement of rounded zeros in compositional data: classical and robust approaches. Comput Stat Data Anal 2012;56:2688e704.
- Martín-Fernández J, Barceló-Vidal C, Pawlowsky-Glahn V. Dealing with zeros and missing values in compositional data sets using nonparametric imputa-tion. Math Geol 2003;35:253e78.
- Aitchison J, Kay JW. Possible solutions of some essential zero problems in compositional data analysis. Compos Data Anal Work Girona 2003; 2003:6.
- Zhao, N., Chen, J., Carroll, I.M., Ringel-Kulka, T., Epstein, M.P., Zhou, H., Zhou, J.J., Ringel, Y., Li, H., and Wu, M.C. (2015) Testing in microbiome-profiling studies with MiRKAT, the microbiome regression-based kernel association test. Am. J. Hum. Genet. 96(5): 797-807.
- Pan, W., Kim, J., Zhang, Y., Shen, X., and Wei, P. (2014) A powerful and adaptive association test for rare variants. *Genetics* 197(4): 1081-95.
- *Cho* I, *Blaser* MJ (*2012*) The human microbiome: at the interface of health and disease. Nat Rev Genet 13: 260–270. doi: 10.1038/nrg3182.
- Li H (2015): Microbiome, Metagenomics and High Dimensional Compositional Data Analysis. *Annual Review of Statistics and Its Application*, 2:73-94.
- Tsilimigras MC, Fodor AA.2016 Compositional data analysis of the microbiome: fundamentals, tools, and challenges. Ann Epidemiol. 2016 May;26(5):330-5.
- Tyler AD, Smith MI, Silverberg MS. Analyzing the human microbiome: a "how to" guide for physicians. Am J Gastroenterol. 2014 Jul;109(7):983-93. doi: 10.1038/ajg.2014.73. Review.
- Blaser MJ. 2014. Missing Microbes. Henry Holt.
- Faust K, Raes J (2012) Microbial interactions: from networks to models. Nature Reviews.
- Weiss et al. (2017) Normalization and microbial differential abundance strategies depend upon data characteristics. Microbiome

References

- Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS, Huttenhower C: **Metagenomic biomarker discovery and explanation.** *Genome Biol* 2011, **12:**R60.
- Paulson JN, Stine OC, Bravo HC, Pop M: **Differential abundance analysis for microbial marker-gene surveys.** *Nat Methods* 2013, **10:**1200-1202.
- Koh H, Blaser MJ, Li H: **A powerful microbiome-based association test and a microbial taxa discovery framework for comprehensive association mapping.** *Microbiome* 2017, **5:**45.
- Benjamini Y, Hochberg Y: **Controlling the false discovery rate: a practical and powerful approach to multiple testing.** *J R Stat Soc Series B Stat Methodol* 1995**:**289-300.
- Yekutieli D: **Hierarchical false discovery rate–controlling methodology.** *J Amer Statistical Assoc* 2008, **103:**309-316.
- Benjamini Y, Yekutieli D: **The false discovery rate approach to quantitative trait loci analysis.** *Genetics* 2005, **171:**783-789.
- Zhang H, DiBaise JK, Zuccolo A, Kudrna D, Braidotti M, Yu Y, Parameswaran P, Crowell MD, Wing R, Rittmann BE: **Human gut microbiota in obesity and after gastric bypass.** *Proc Natl Acad Sci U S A* 2009, **106:**2365-2370.
- Collado MC, Isolauri E, Laitinen K, Salminen S: **Distinct composition of gut microbiota during pregnancy in overweight and normal-weight women.** *Am J Clin Nutr* 2008, **88:**894-899.