

18 Jan. 2024

BIBS Seminar

METHOD

Open Access



Population structure discovery in meta-analyzed microbial communities and inflammatory bowel disease using MMUPHin

Siyuan Ma¹, Dmitry Shungin², Himel Mallick², Melanie Schirmer², Long H. Nguyen³, Raivo Kolde², Eric Franzosa¹, Hera Vlamakis², Ramnik Xavier^{2*} and Curtis Huttenhower^{1*} 

Presenter : Seungrin Yang

Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul, Korea

Contact : ysl.bibs.snu@gmail.com

Keywords : Inflammatory bowel disease, Metagenomics, Dysbiosis, Meta-analysis, Batch effect

Outline

- Introduction
- Method
- Materials
- Results
- Application to an external dataset
- Conclusion

Outline

- Introduction
- Method
- Materials
- Results
- Application to an external dataset
- Conclusion

Meta-analysis for molecular epidemiology in large populations has seen great success in linking high-dimensional 'omic features to complex health-related phenotypes.

One example of this is in Genome-wide association studies (GWAS [1]), where appropriate study scale, **achieved by rigorous integration of multiple cohorts**, has both facilitated reproducible discoveries (in the form of disease-associated loci[2-4]) and addressed **confounding due to unobserved population structure** [5].



Inflammatory bowel disease (IBD, 염증성 장질환) represent a particular success story for GWAS **meta-analysis**.

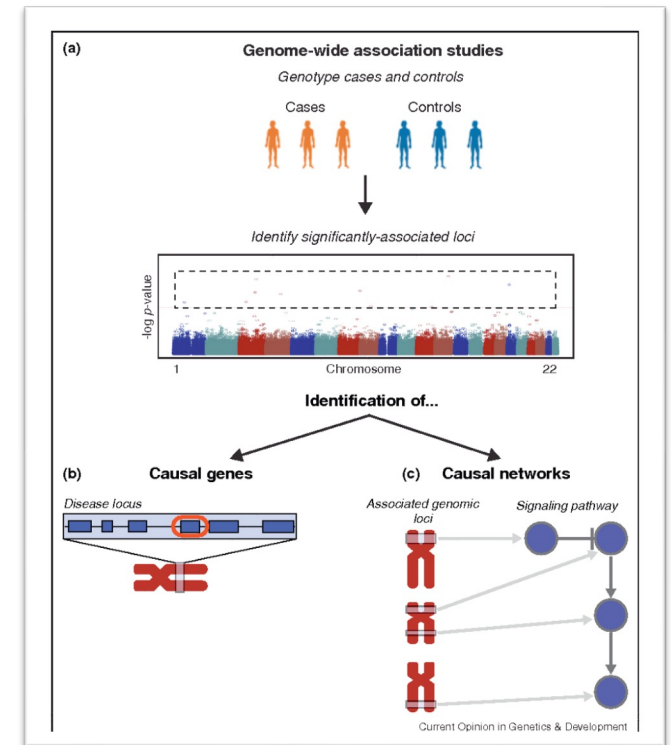



Image source: [10.1016/j.jde.2013.09.003](https://doi.org/10.1016/j.jde.2013.09.003)

Introduction

 Recap: GWAS aims to find the associations between genetic variations and observable trait

What's the problem?

Meta-analysis of microbial community profiles presents unique quantitative challenges relative to other types of 'omics data such as GWAS [10] or gene expression [11].

These include particularly strong batch, inter-individual, and inter-population differences, and statistical issues including zero-inflation and compositionality [12, 13].

Consequently, methods to correct for cohort and batch effects from other 'omics settings [14–17] are not directly appropriate

Introduction

What's the problem?

To date, there are no methods permitting the joint analysis of batch-corrected microbial profiles for most study designs.

In the absence of methods appropriate for large-scale microbial meta-analysis, it is unclear whether reproducible population structure in the microbiome, such as microbially driven IBD “subtypes”, exists to help explain the clinical heterogeneity of these conditions.

In this work, a uniform statistical framework for population-scale meta-analysis of microbiome data is introduced and validated

Introduction

A statistical framework for meta-analysis of microbial community profiles

MMUPHIn, A collection of novel methods for **meta-analysis** of environmental exposures, phenotypes, and **population structures** across microbial community studies, specifically accounting for technical batch effects and interstudy differences (“Methods,” Fig. 1a).

💡 Meta-analysis?

Meta-analysis in this context likely involves synthesizing data from multiple studies to gain a more comprehensive understanding of the microbial population structure.

💡 Population structure :

This term generally refers to the composition of a community in terms of its constituent members. It means the different types of microorganisms present in a community, their relative abundances, and how they are organized or distributed.

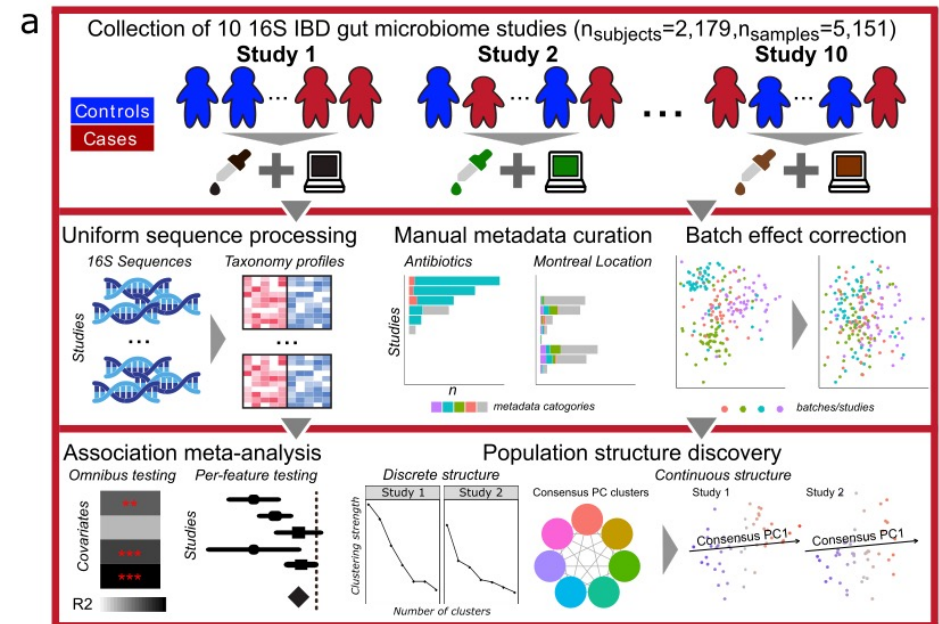


Fig 1. A method for large-scale microbial community meta-analysis and its application to inflammatory bowel disease

Introduction

Outline

- Introduction
- **Method**
- Data
- Results
- Application to an external dataset
- Conclusion

MMUPHIn_Correct – Batch and Study effect correction

MMUPHIn (*Meta-Analysis Methods with a Uniform Pipeline for Heterogeneity in* microbiome studies)

For microbial community batch correction, they extended the batch correction method developed for gene expression data in ComBat[15] with an additional component to allow for the zero-inflated nature of microbial abundance data.

Let's say sample read count Y was modeled with respect to both batch variable and biologically relevant covariate(s) X :

$$Y_{ijp} = \exp\{\beta_p X'_{ij} + \sigma(\gamma_{ip} + \delta_{ip} \epsilon_{ijp})\} \times I_{ijp} ,$$

where

- i indicates batch/study
- j indicates sample
- p indicates feature

Methods

MMUPHin_Correct – Batch and Study effect correction

Let's break down

The linear part of the model, $\beta_p X'_{ij} + \sigma(\gamma_{ip} + \delta_{ip} \epsilon_{ijp})$, is the combination of effects from covariates, batch variable, random error.

However, this linear combination can result in any real number, including negatives.

$$Y_{ijp} = \exp\left\{ \underbrace{\beta_p X'_{ij}}_{\substack{1) \\ \text{covariates}}} + \sigma \left(\underbrace{\gamma_{ip}}_{\substack{2) \\ \text{batch variables}}} + \underbrace{\delta_{ip} \epsilon_{ijp}}_{\substack{3) \\ \text{random error}}} \right) \right\} \times I_{ijp},$$

Methods

MMUPHin_Correct – Batch and Study effect correction

Let's break down

The linear part of the model, $\beta_p X'_{ij} + \sigma(\gamma_{ip} + \delta_{ip} \epsilon_{ijp})$, is the combination of effects from covariates, batch variable, random error.

However, this linear combination can result in any real number, including negatives.

$$Y_{ijp} = \exp\left\{ \underbrace{\beta_p X'_{ij}}_{\substack{1) \\ \text{covariates}}} + \underbrace{\sigma(\gamma_{ip} + \delta_{ip} \epsilon_{ijp})}_{\substack{2) \\ \text{batch variables} \quad \text{3) \\ \text{random error}}} \right\} \times I_{ijp}$$

To ensure that the outcome (Y_{ijp}) remains positive and interpretable in the context of count data, the linear predictor is placed inside an exponential function.

By using the exponential of a linear combination of variables, the model implicitly assumes that the count data follow a log-normal distribution, a common assumption for many types of biological data.

Methods

MMUPHin_Correct – Batch and Study effect correction

Let's break down

$$Y_{ijp} = \exp\{\beta_p X'_{ij} + \sigma(\gamma_{ip} + \delta_{ip} \epsilon_{ijp})\} \times I_{ijp},$$

A feature-specific standardization factor

Covariate-specific coefficients

Batch-specific location and scale parameters

Independent error term following standard normal distribution.

$\epsilon_{ijp} \sim N(0,1)$

Methods

MMUPHin_Correct – Batch and Study effect correction

Let's break down

Modelled with normal prior
 $\gamma_{ip} \sim N(Y_i, \tau_i^2)$

Modelled with inverse-gamma prior
 $\delta_{ip} \sim \text{Inverse Gamma}(\lambda_i, \theta_i)$

$$Y_{ijp} = \exp\{\beta_p X'_{ij} + \sigma(\underline{\gamma_{ip}} + \underline{\delta_{ip}} \epsilon_{ijp})\} \times \underline{I_{ijp}},$$

A binary (0,1) zero-count indicator, to allow for zero-inflation of features.

Hyperparameters $(Y_i, \tau_i^2, \lambda_i, \theta_i)$ are estimated with empirical Bayes estimators as in ComBat

Methods

MMUPHin_Correct – Batch and Study effect correction

The posterior means, γ_{ip}^* and δ_{ip}^* , along with standard frequentist estimates $\hat{\beta}_p$ and $\hat{\sigma}_p$ are used to provide batch-corrected count data:

$$\widetilde{Y}_{ijp} = \exp \left\{ \frac{Y_{ijp} - \hat{\beta}_p X'_{ij} - \gamma_{ip}^* \hat{\sigma}_p}{\delta_{ip}^*} + \hat{\beta}_p X'_{ij} \right\} \times I_{ijp} ,$$

where

- i indicates batch/study
- j indicates sample
- p indicates feature

Methods

MMUPHin_Correct – Batch and Study effect correction

Let's break down

Original sample read count data

$$Y_{ijp} = \exp\{\beta_p X'_{ij} + \sigma(\gamma_{ip} + \delta_{ip}\epsilon_{ijp})\} \times I_{ijp}$$

The batch effect that needs to be adjusted.

By subtracting this from the original read count Y_{ijp} , the formula aims to remove the influence of batch-specific variations

$$\widetilde{Y}_{ijp} = \exp \left\{ \frac{Y_{ijp} - \hat{\beta}_p X'_{ij} - \gamma_{ip}^* \hat{\sigma}_p}{\delta_{ip}^*} + \hat{\beta}_p X'_{ij} \right\} \times I_{ijp},$$

batch-specific scale parameter δ_{ip}^*

The difference $(Y_{ijp} - \hat{\beta}_p X'_{ij} - \gamma_{ip}^* \hat{\sigma}_p)$ is then scaled by the batch-specific scale parameter δ_{ip}^* .

This scaling adjusts the variance of the read counts, aligning them across different batches.

the covariate effects

After adjusting for batch effects and scaling, the covariate effects $\hat{\beta}_p X'_{ij}$ are added back. This ensures that the biological information encoded in the covariates is retained in the batch-corrected data.

Methods

MMUPHin_Correct – Batch and Study effect correction

In practice, the user provides

- sample microbial abundance table(Y),
- batch/study information, and
- optionally any other covariates X that are potentially confounded with batch but encode important biological information.

Example R codes)

```
fit_adjust_batch <- adjust_batch(feature_abd = CRC_abd,  
                                batch = "studyID",  
                                covariates = "study_condition",  
                                data = CRC_meta,  
                                control = list(verbose = FALSE))  
  
CRC_abd_adj <- fit_adjust_batch$feature_abd_adj
```

MMUPHin software [36] is available at Bioconductor.

Methods

Outline

- Introduction
- Method
- **Materials**
- Results
- Application to an external dataset
- Conclusion

MMUPHin_Correct – Batch and Study effect correction

With this model specification, MMUPHin_Correct is expected to often reduce, rather than fully correct batch differences.

This is because MMUPHin_Correct focuses on correcting non-zero abundance batch effects, and does not change features' presence/absence across batches.

$$\widetilde{Y}_{ijp} = \exp \left\{ \frac{Y_{ijp} - \hat{\beta}_p X'_{ij} - \gamma_{ip}^* \hat{\sigma}_p}{\delta_{ip}^*} + \hat{\beta}_p X'_{ij} \right\} \times \underline{I_{ijp}}$$

A binary (0,1) zero-count indicator, to allow for zero-inflation of features.

Materials

MMUPHin_Correct – Batch and Study effect correction

- “Correcting” a feature’s batch-specific presence to absence is inappropriate, as substantial non-zero read counts indicate biological presence rather than technical artifacts.
- Imputing non-zero abundance for batch-specific absence is technically challenging in our linear modelling framework, as the per-sample/feature noise ϵ_{ijp} cannot be reliably inferred for inflated zero values.

Ten uniformly processed 16S rRNA gene sequencing studies of the IBD mucosal/stool microbiomes were used

- They collected and uniformly processed ten published 16S studies of the IBD gut microbiome totaling 2179 subjects and 5151 samples.
- These studies range widely in terms of cohort designs and population characteristics, including
 - recent-onset and established disease patients,
 - cross-sectional and longitudinal sampling,
 - pediatric and adult populations,
 - diseases (CD and UC),
 - treated and treatment-naive patients,
 - biopsy and stool samples, and
 - inclusion of healthy/non- IBD controls.
- Covariates were manually curated to ensure consistency across studies (“[Methods](#)”).
- Major factors available from all or most studies included
 - demographics (age/sex/race), biogeography, disease location and/or extent, antibiotic usage, immunosuppression, and steroid and/or 5-ASA usage.

Materials

Ten uniformly processed 16S rRNA gene sequencing studies of the IBD mucosal/stool microbiomes were used

Study	Brief description	N subject	N sample	Phenotype(s)	Age	Gender	Sample type(s)									
PROTECT [23]	Longitudinal cohort of newly diagnosed UC	405	1212 (539)	UC 405	12.71 (3.29)	Male 52%/ Female 48%	Biopsy 22%/ Stool 78%	Jansson-Lamendella [22]	Longitudinal follow-up with fecal samples	137	683 (137)	CD 49/UC 60/ Control 28		Male 42%/ Female 58%	Stool	
RISK [7]	Pediatric cohort of treatment-naïve CD	631	882	CD 430/Control 201	12.16 (3.22)	Male 59%/ Female 41%	Biopsy 72%/ Stool 28%	Pouchitis [27]	Patients recruited underwent IPAA for treatment of UC or FAP prior to enrollment.	353	577	CD 42/UC 266/ Control 45	46.19 (13.58)	Male 52%/ Female 48%	Biopsy	
Herfarth [26]	Densely (daily) sampled	31	860 (31)	CD 19/Control 12	36.03 (14.12)	Male 35%/ Female 58%/Miss-	Stool									
								CS-PRISM [28]	Cross-sectional cohort nested in PRISM	397	467	CD 215/UC 144/Control 38	41.68 (15.22)	Male 47%/ Female 53%	Biopsy 29%/ Stool 71%	
								HMP2 [9]	Large cohort of newly diagnosed IBD with multi-omics measurement.	81	177 (162)	CD 37/UC 22/ Control 22	29.76 (19.63)	Male 51%/ Female 49%	Biopsy	
								MucosalIBD [29]	Pediatric cohort with Paneth cell phenotypes	83	132	CD 36/Control 47	12.93 (3.65)	Male 58%/ Female 42%	Biopsy	
								LSS-PRISM [30]	Longitudinal cohort nested in PRISM.	18	88 (19)	CD 12/UC 6	30.37 (10.52)	Male 39%/ Female 61%	Stool	
								BIDMC-FMT [31]	FMT Trial design	8	16	CD 8	38.38 (12.73)	Male 62%/ Female 38%	Stool	

(left , right) Table 1. Ten uniformly processed 16S rRNA gene sequencing studies of the IBD mucosal/stool microbiomes.

For longitudinal cohorts, numbers in parentheses indicate baseline sample size.

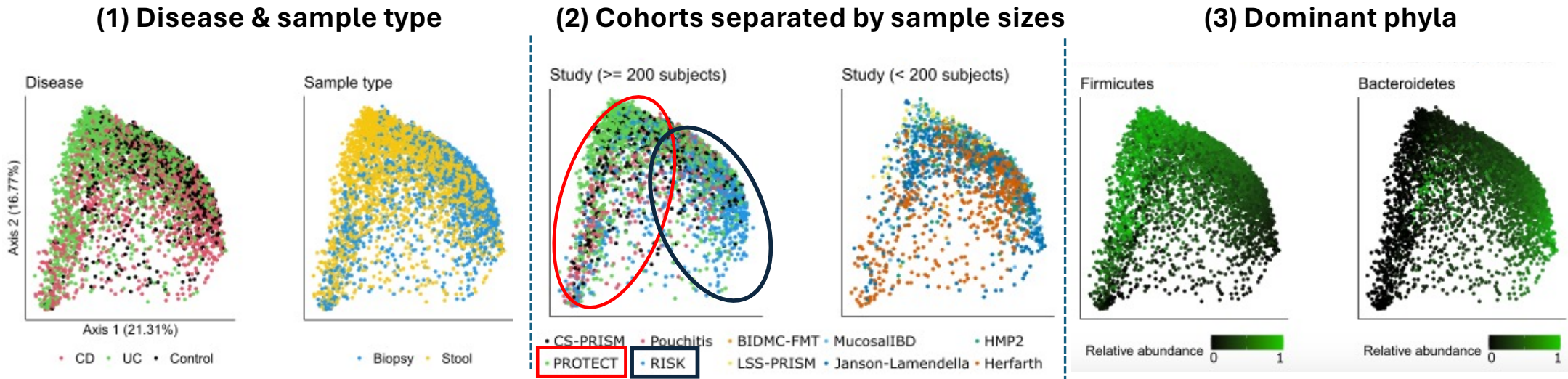
For age, mean and standard error (parenthesized) are shown. Additional covariates are summarized in

Additional file 3: Table S1

Materials

Ten uniformly processed 16S rRNA gene sequencing studies of the IBD mucosal/stool microbiomes were used

Fig 1b. MDS ordination of all microbial profiles (Bray-Curtis dissimilarity) before batch correction visualize the strongest associations with gut microbial composition, including disease, sample type (biopsy or stool), cohort (visualized separately for larger and smaller studies), and dominant phyla

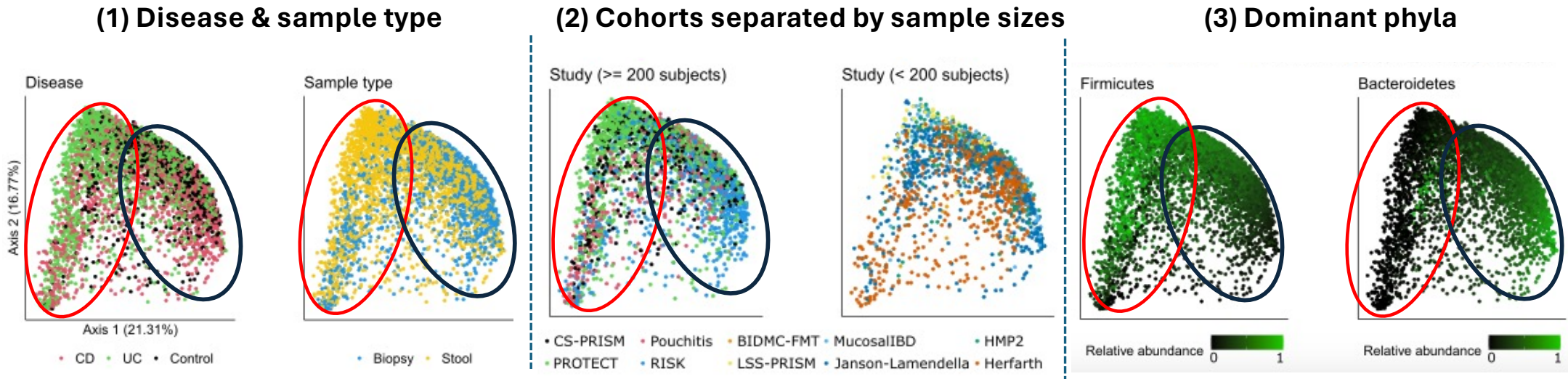


- Microbiome differences associated with disease were notable even without normalization.
- However, this can be misleading due to the confounding of cohort structure between studies, such as the differentiation between **RISK** (a predominantly mucosal study of CD) and **PROTECT** (a predominantly stool study of UC).

Materials

Ten uniformly processed 16S rRNA gene sequencing studies of the IBD mucosal/stool microbiomes were used

Fig 1b. MDS ordination of all microbial profiles (Bray-Curtis dissimilarity) before batch correction visualize the strongest associations with gut microbial composition, including disease, sample type (biopsy or stool), cohort (visualized separately for larger and smaller studies), and dominant phyla



- Inter-individual differences largely independent of population or disease, such as Bacteroidetes versus Firmicutes dominance, were also universal among studies and sample types as expected [9, 32].

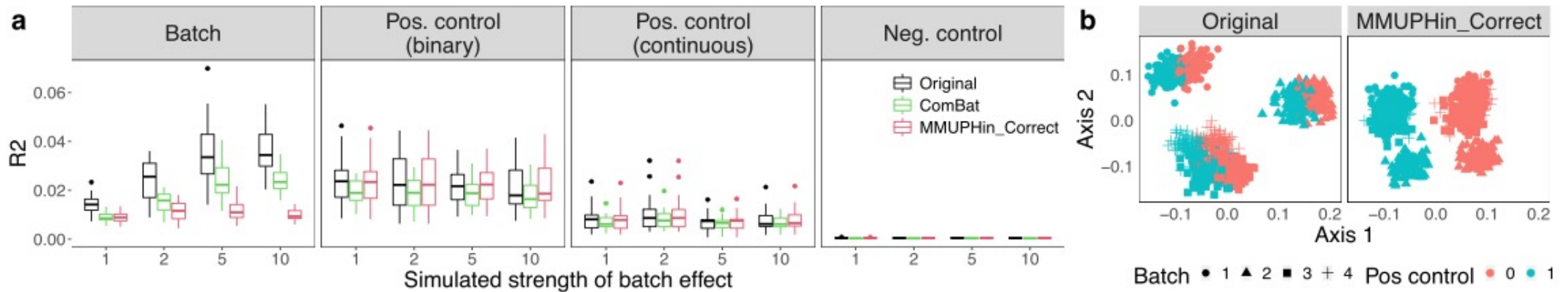
Materials

Outline

- Introduction
- Method
- Materials
- **Results**
- Application to an external dataset
- Conclusion

Ten uniformly processed 16S rRNA gene sequencing studies of the IBD mucosal/stool microbiomes were used

Fig 2a, b. MMUPHIn_Correct is effective for covariate-adjusted batch effect reduction while maintaining the effect of positive control variables



- For panel a, PERMANOVA R2 statistics summarize the effect of batch and positive/negative control variables on the overall microbial composition, before and after batch correction.

Results

Outline

- Introduction
- Method
- Materials
- Results
- **Application to an external dataset**
- Conclusion

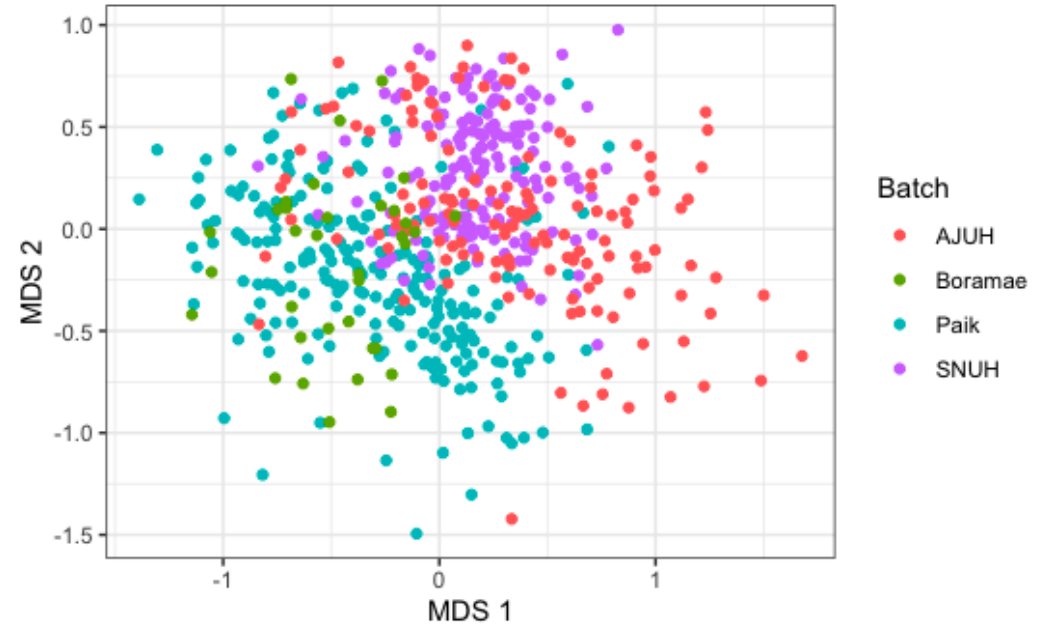
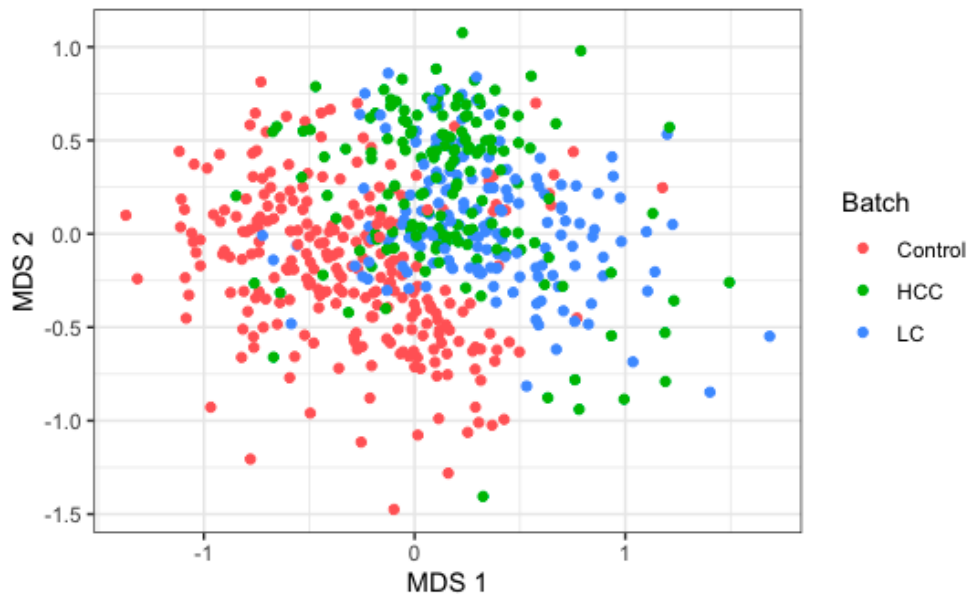
Datasets

Liver cancer serum dataset	
Approach	Metagenome
Sample size	571
No. of OTUs	547
No. of batches	4
Batch effect source	Institutions (SNUH,AJUH,Paik, Borame)
Sample type	Serum sample
Study design	Cross-section study
Sequence region	V3-V4

Application to an external dataset

Application to liver cancer serum dataset

MDS plots

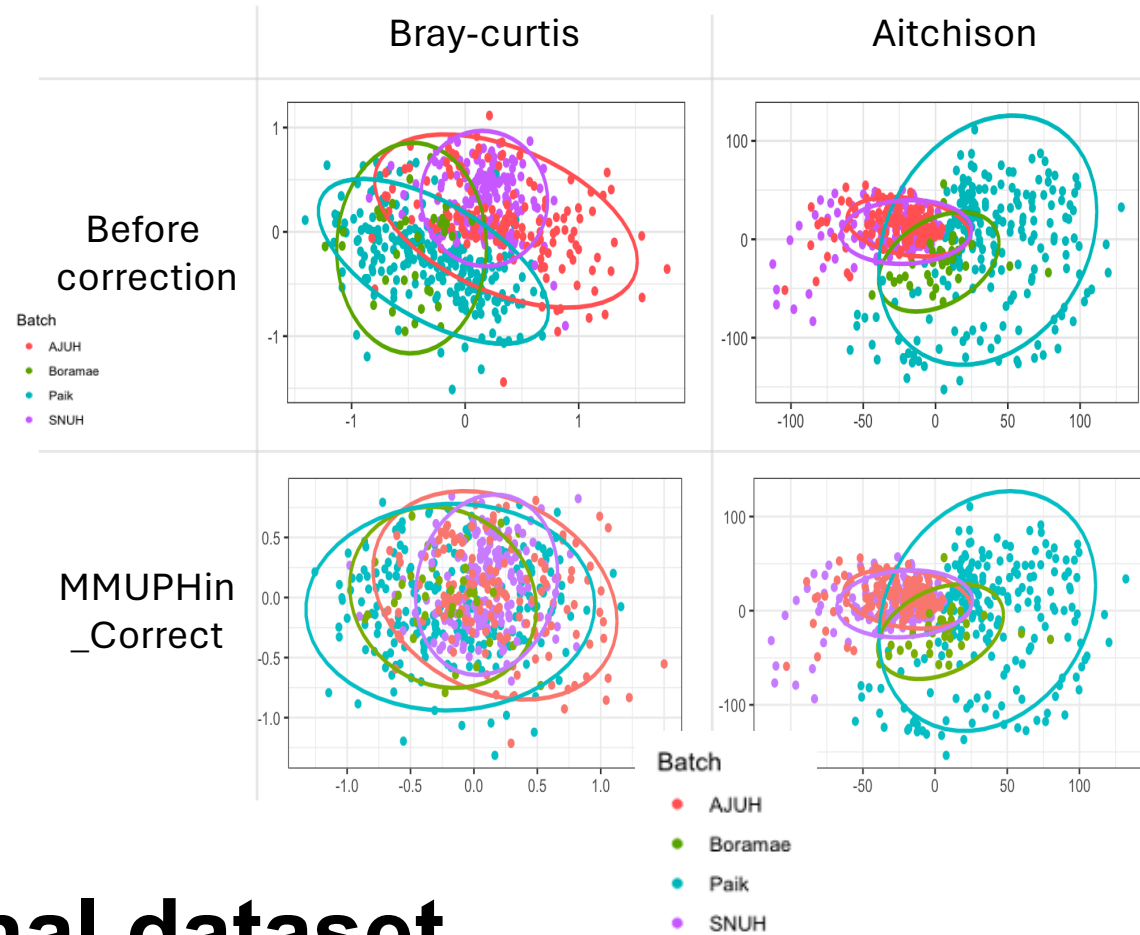


Application to an external dataset

Application to liver cancer serum dataset

Results) PERMANOVA R^2 (left) and MDS plot (right)

Methods	Dissimilarity	
	Bray-Curtis	Aitchison
Before correction	0.0992	0.0899
Percentile-Normalization	0.1812652	0.09898626
ConQur <small>(Ling et al., 2022)</small>	0.0699	0.3336
MMUPHIn	0.0508	0.0858



Application to an external dataset

Outline

- Introduction
- Method
- Materials
- Results
- Application to an external dataset
- **Conclusion**

A novel framework for microbial community meta-analysis

- The meta-analysis framework developed for the study, MMUPHIn, has been extensively evaluated and its performance for batch effect removal, supervised meta-analysis of exposures and covariates, and unsupervised pollution structure discovery validated on a variety of simulated microbial community types.
- In this seminar, we focused on the batch effect removal method, MMUPHIn_Correct, and applied them to external dataset (liver cancer dataset).
- It shows moderate adjustment of the batch effects from its own dataset and as well as the external dataset.
- While it is extended version of ComBat to microbiome analysis by considering zero-inflation, it assumes the data to be zero-inflated Gaussian, which is only appropriate for certain transformations of relative abundance data (i.e., taxon counts normalized by each sample's library size). Therefore, more flexible approaches are needed (Ling et al., 2022)

Conclusion

Thank you