Differential Abundance Analysis of Shotgun Metagenomic Sequencing Data

Carol Wang

Duke University

December 15, 2022

Shotgun metagenomic sequencing

- Sequence all microbial genomic DNA in the sample
- Richer information about the microbial composition and gene functions; Detection of new species and new genes
- Questions
  - Who's there? (Taxonomic profile)
  - What they are doing? (Functional profile)
- Integrated preprocessing tools: Biobakery 3 (Beghini et al., 2021)
- Preprocessing pipelines for taxonomic profiling
  - Alignment-based, e.g., MetaPhlAn (**?**)
  - Composition-based

Taxonomic profiling via shotgun metagenomic sequencing

- A typical dataset
  - Taxa abundance
  - Phylogenetic tree
  - Design features
    - Longitudinal sampling
    - Contrasting groups: case control
    - Body sites
    - Covariates: sample-level and/or subject-level
- Question of interest: Differential abundance analysis of microbiome compositions

Two types of abundance

- Counts
- Observed relative abundance

Modeling the counts

- Why using counts?

  Suppose there are three species A,B,C. $(n_A, n_B, n_C) = (10, 10, 80)$ is more informative than $(n_A, n_B, n_C) = (1, 1, 8)$ though the observed relative abundance is the same.

Modeling the counts: Setup

- Data: read counts $\boldsymbol{X} = \{X_{ik}\}$, where X_{ik} is the number of aligned reads to marker gene k in sample i
- Each species j has its unique set of marker genes S_j. In the table below, $S_1 = \{A, B, C\}, S_2 = \{D, E\}$.
- How to deal with the read counts of the multiple marker genes?

  Key assumption: Marker effects on read counts can be approximated with marker length.

	species 1			species 2	
	marker A	marker B	marker C	marker D	marker E
sample 1	X_{1A}	X_{1B}			
sample 2	X_{2A}		...		
sample 3	X_{3A}				

The MetaPhlAn estimate

Let X_{ik} be number of reads mapped to marker k in sample i, l_k the length of marker k, S_j the set of markers of species j, θ_{ij} the underlying relative abundance of species j in sample i.

- The MetaPhlAn estimate (Segata et al.,2012)

$$\hat{\theta}_{ij} \propto \frac{\sum_{k \in S_j} X_{ik}}{\sum_{k \in S_j} l_k}$$

- The implied multinomial sampling model of clades

$$\{N_{ij}\}_{j=1}^{J} | N_i, \boldsymbol{p_i} \overset{\text{ind}}{\sim} \text{Multinomial}(N_i, \boldsymbol{p_i}),$$

where $N_{ij} = \sum_{k \in S_j} X_{ik}$, $N_i = \sum_{j=1}^{J} N_{ij}$ and $p_{ij} = \frac{\theta_{ij} \sum_{k \in S_j} l_k}{\sum_{j=1}^{J} \theta_{ij} \sum_{k \in S_j} l_k}$

- Multiplicative effect from marker length

MetaPhlAn: perturbation

Definition 1. The *constraining operator* $\mathcal{C}$ transforms a vector $\boldsymbol{w}$ of D non-negative components and positive sum into the unit-sum vector $\frac{\boldsymbol{w}}{\sum_{j=1}^{D} w_j}$.

Definition 2. Let $\boldsymbol{w}$ be a D-part composition and $\boldsymbol{u}$ a D-vector with positive elements. Then the operation

$$\boldsymbol{W} = \boldsymbol{u} \circ \boldsymbol{w} = \mathcal{C}(u_1 w_1, \cdots, u_D w_D)$$

is termed a *perturbation* with the original composition $\boldsymbol{w}$ being operated on by the *perturbing vector* $\boldsymbol{u}$ to form a *perturbed composition* $\boldsymbol{W}$.

- The multinomial parameter $\boldsymbol{p}_i$ in MetaPhlAn is a perturbed composition

$$\boldsymbol{p}_i = \boldsymbol{u} \circ \boldsymbol{\theta}_i$$

where $u_j = \sum_{k \in S_j} l_k$.

MSSQ

- MSSQ (**?**)

$$X_{ik} \overset{\text{ind}}{\sim} \text{Poisson}(\theta_{ij} t_i \phi_k l_k)$$

  where ϕ_k is the marker gene specific effect, t_i the total (mapped) reads in sample i, and $\sum_j \theta_{ij} = 1$.

- The implied multinomial sampling model

$$\{N_{ij}\}_{j=1}^{J} | N_i, \boldsymbol{p_i}(\boldsymbol{\phi}) \overset{\text{ind}}{\sim} \text{Multinomial}(N_i, \boldsymbol{p_i}(\boldsymbol{\phi})),$$

  where

$$\boldsymbol{p_i}(\boldsymbol{\phi}) = \boldsymbol{u}(\boldsymbol{\phi}) \circ \boldsymbol{\theta_i}$$

$$\boldsymbol{u}(\boldsymbol{\phi}) = \{\sum_{k \in S_j} \phi_k l_k : j = 1, \cdots, J\}$$

- Setting $\phi_k = 1$, the MLE of $\boldsymbol{\theta}$ is the MetaPhlAn estimate
- For low abundance species, ϕ_k is hard to estimate

Recap: LTN for 16S data

- Multinomial sampling model of the counts

$$\boldsymbol{X}_i|\boldsymbol{p}_i, N_i \overset{\text{ind}}{\sim} \text{Multinomial}(N_i, \boldsymbol{p}_i)$$

- *Tree-based logratio* (tlr) transform of $\boldsymbol{p}_i$

$$tlr_{\mathcal{T}}(\boldsymbol{p}_i) = \left\{ \log\left(\frac{\sum_{j \in A_l} p_{ij}}{\sum_{j \in A_r} p_{ij}} \right) : A \in \mathcal{T} \right\}$$

- Gaussian model for tree-based logratios

$$tlr(\boldsymbol{p}_i) \overset{\text{iid}}{\sim} \text{MVN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

Perturbation along the tree

- Perturbation is equivalent to addition of additive logratios

$$\boldsymbol{p}_i = \boldsymbol{u} \circ \boldsymbol{\theta}_i \Leftrightarrow alr(\boldsymbol{p}_i) = alr(\mathcal{C}(\boldsymbol{u})) + alr(\boldsymbol{\theta}_i)$$

  where $alr(\boldsymbol{x}) = \{\log\left(\frac{x_j}{x_D}\right) : j = 1, \cdots, D-1\}$ for a D-composition $\boldsymbol{x}$.

- Marker effects $\boldsymbol{u}$ can be decomposed along the tree like the composition $\boldsymbol{\theta}_i$ due to symmetry
- An alternative sampling model based on MetaPhlAn assumptions

$$N_i(A_l)|N_i(A), q_i(A) \overset{\text{ind}}{\sim} \text{Binomial}(N_i(A), q_i(A))$$

  where $N_i(A) = \sum_{j \in A} N_{ij}$ and

$$\frac{q_i(A)}{1 - q_i(A)} = \frac{\sum_{j \in A_l} \theta_{ij}}{\sum_{j \in A_r} \theta_{ij}} \cdot \frac{\sum_{j \in A_l} u_j}{\sum_{j \in A_r} u_j}$$

- Addition of tree-based logratios

$$tlr(\boldsymbol{p}_i) = tlr(\boldsymbol{u}) + tlr(\boldsymbol{\theta}_i)$$

Model for n exchangeable samples

- Binomial sampling on the tree

$$N_i(A_l)|N_i(A), q_i(A) \overset{\text{ind}}{\sim} \text{Binomial}(N_i(A), q_i(A)) \qquad i = 1, \cdots, n, A \in \mathcal{T}$$

- Marker effects on tree-based logratios

$$\log \frac{q_i(A)}{1 - q_i(A)} = \psi_i(A) + v(A) \qquad i = 1, \cdots, n, A \in \mathcal{T}$$

where $\psi_i(A) = \log \frac{\sum_{j \in A_l} \theta_{ij}}{\sum_{j \in A_r} \theta_{ij}}$, $v(A) = \log \frac{\sum_{j \in A_l} u_j}{\sum_{j \in A_r} u_j}$ accounts for the marker effect due to their difference in length

- Latent Gaussian representation of species compositions

$$\boldsymbol{\psi}_i \overset{\text{iid}}{\sim} \text{MVN}(\boldsymbol{\mu}, \boldsymbol{\Sigma}_\psi)$$

Mixed-effects modeling

- Subgrouping structure of the samples can be characterized with mixed-effects model on the log-odds:

$$\{N_{ij}\}_{j=1}^{J} | N_i, \boldsymbol{p_i} \overset{ind}{\sim} \text{Multinomial}(N_i, \boldsymbol{p_i})$$
$$tlr_{\mathcal{T}}(\boldsymbol{p}_i) = tlr_{\mathcal{T}}(\boldsymbol{u}) + s_i\boldsymbol{\alpha} + \boldsymbol{\beta}\boldsymbol{Z}_i + \boldsymbol{\gamma}_{g_i} + \boldsymbol{\epsilon}_i$$
$$\boldsymbol{\gamma}_{g_i} \overset{iid}{\sim} MVN(\boldsymbol{\mu}, \boldsymbol{\Omega}^{-1})$$
$$\boldsymbol{\epsilon_i} \overset{iid}{\sim} MVN(\mathbf{0}, diag(\sigma_1^2, \cdots, \sigma_{J-1}^2))$$

  where $\boldsymbol{Z}_i$ are the covariates of sample i, g_i is the subgroup that sample i belongs to, s_i is the univariate variable and we aim to test whether the microbiome composition differ between samples with different values of s_i.

- Prior specification
  - Gaussian prior on $\boldsymbol{\mu}, \boldsymbol{\beta}$
  - Graphical Lasso prior on $\boldsymbol{\Omega}$
  - Spike-and-slab prior on $\boldsymbol{\alpha}$ for testing purpose.

Data processing

Data processing

- We processed metagenomic sequencing data from The Inflammatory Bowel Disease Multi'omics Database with MetaPhlAn
- The raw sequences and the (relative) taxa abundance is available here
- Focus on the 760 samples with MGX-MTX pairs

Data processing: from FASTQ to counts

Step 1. For each raw sequence saved in FASTQ file, run MetaPhlAn to obtain marker counts

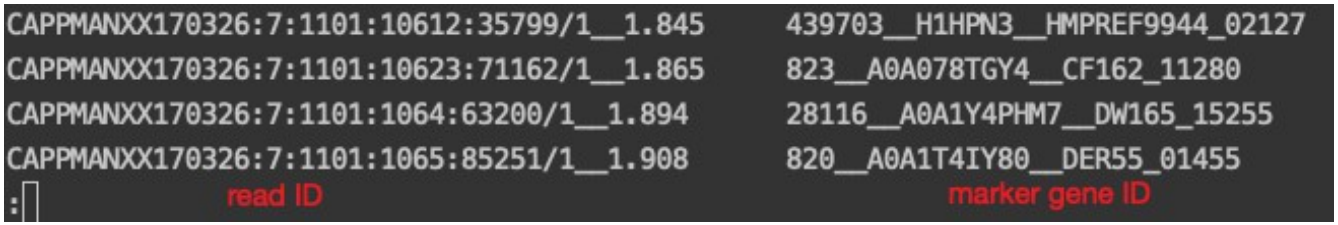
```
MetaPhlAn HSM7J4PQ.fastq.gz --bowtie2out HSM7J4PQ.bowtie2.bz2
--input_type fastq -o profiled_HSM7J4PQ.txt
```

```
MetaPhlAn -t marker_counts HSM7J4PQ.bowtie2.bz2 --input_type bowtie2out
-o marker_counts_HSM7J4PQ.txt
```

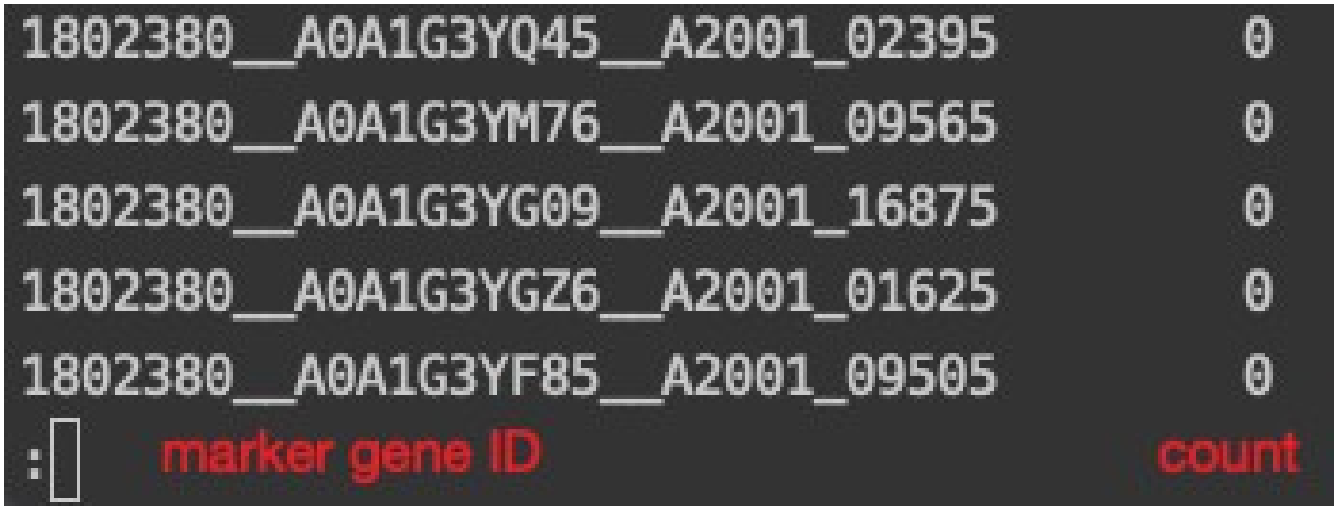
Here `-t marker_counts` specifies the analysis type as the marker counts.

Outputs

- `HSM7J4PQ.bowtie2.bz2`: mapping results of reads to marker genes



- `marker_counts_HSM7J4PQ.txt`: marker counts $\boldsymbol{X} = (X_{ik})$



The count table

Step 2. Merge the marker counts into a count table.
The resulting count table has 757 samples and 68829 markers.

Taxonomic information

Step 3. Assign taxonomy to the markers using the marker information available here. The 68829 markers belong to 2227 species.

Marker information

```
r$> head(marker_info)
  X                          gene
1 0 1802380__A0A1G3YJF8__A2001_04380
2 1 1802380__A0A1G3YGX7__A2001_01635
3 2 1802380__A0A1G3YPL7__A2001_15105
4 3 1802380__A0A1G3YGU4__A2001_01425
5 4 1802380__A0A1G3YIS5__A2001_03140
6 5 1802380__A0A1G3YU64__A2001_12365

                                        dict_str
1  {'clade': 's__Treponema_sp_GWC1_61_84', 'ext': [], 'len': 645, 'score': 0, 'taxon': 'k__Bacteria|p__Spirochaetes|c__Spirochaetia|o__Spirochaetales|f__Spir
ochaetaceae|g__Treponema|s__Treponema_sp_GWC1_61_84'}
2  {'clade': 's__Treponema_sp_GWC1_61_84', 'ext': [], 'len': 894, 'score': 0, 'taxon': 'k__Bacteria|p__Spirochaetes|c__Spirochaetia|o__Spirochaetales|f__Spir
ochaetaceae|g__Treponema|s__Treponema_sp_GWC1_61_84'}
3  {'clade': 's__Treponema_sp_GWC1_61_84', 'ext': [], 'len': 510, 'score': 0, 'taxon': 'k__Bacteria|p__Spirochaetes|c__Spirochaetia|o__Spirochaetales|f__Spir
ochaetaceae|g__Treponema|s__Treponema_sp_GWC1_61_84'}
4  {'clade': 's__Treponema_sp_GWC1_61_84', 'ext': [], 'len': 567, 'score': 0, 'taxon': 'k__Bacteria|p__Spirochaetes|c__Spirochaetia|o__Spirochaetales|f__Spir
ochaetaceae|g__Treponema|s__Treponema_sp_GWC1_61_84'}
5 {'clade': 's__Treponema_sp_GWC1_61_84', 'ext': [], 'len': 2901, 'score': 0, 'taxon': 'k__Bacteria|p__Spirochaetes|c__Spirochaetia|o__Spirochaetales|f__Spir
ochaetaceae|g__Treponema|s__Treponema_sp_GWC1_61_84'}
6  {'clade': 's__Treponema_sp_GWC1_61_84', 'ext': [], 'len': 954, 'score': 0, 'taxon': 'k__Bacteria|p__Spirochaetes|c__Spirochaetia|o__Spirochaetales|f__Spir
ochaetaceae|g__Treponema|s__Treponema_sp_GWC1_61_84'}
                      clade ext  len score
1 s__Treponema_sp_GWC1_61_84  []  645     0
2 s__Treponema_sp_GWC1_61_84  []  894     0
3 s__Treponema_sp_GWC1_61_84  []  510     0
4 s__Treponema_sp_GWC1_61_84  []  567     0
5 s__Treponema_sp_GWC1_61_84  [] 2901     0
6 s__Treponema_sp_GWC1_61_84  []  954     0
                                                                                                                                                  taxon
1 k__Bacteria|p__Spirochaetes|c__Spirochaetia|o__Spirochaetales|f__Spirochaetaceae|g__Treponema|s__Treponema_sp_GWC1_61_84
2 k__Bacteria|p__Spirochaetes|c__Spirochaetia|o__Spirochaetales|f__Spirochaetaceae|g__Treponema|s__Treponema_sp_GWC1_61_84
3 k__Bacteria|p__Spirochaetes|c__Spirochaetia|o__Spirochaetales|f__Spirochaetaceae|g__Treponema|s__Treponema_sp_GWC1_61_84
4 k__Bacteria|p__Spirochaetes|c__Spirochaetia|o__Spirochaetales|f__Spirochaetaceae|g__Treponema|s__Treponema_sp_GWC1_61_84
5 k__Bacteria|p__Spirochaetes|c__Spirochaetia|o__Spirochaetales|f__Spirochaetaceae|g__Treponema|s__Treponema_sp_GWC1_61_84
6 k__Bacteria|p__Spirochaetes|c__Spirochaetia|o__Spirochaetales|f__Spirochaetaceae|g__Treponema|s__Treponema_sp_GWC1_61_84
```

Phylogenetic information

Step 4. Obtain the phylogenetic tree for this dataset by pruning the full phylogenetic tree for the MetaPhlAn 3.1 database.
The species (and their markers) that are not leaves of the full tree are removed. $< 7\%$ of the marker counts are removed.

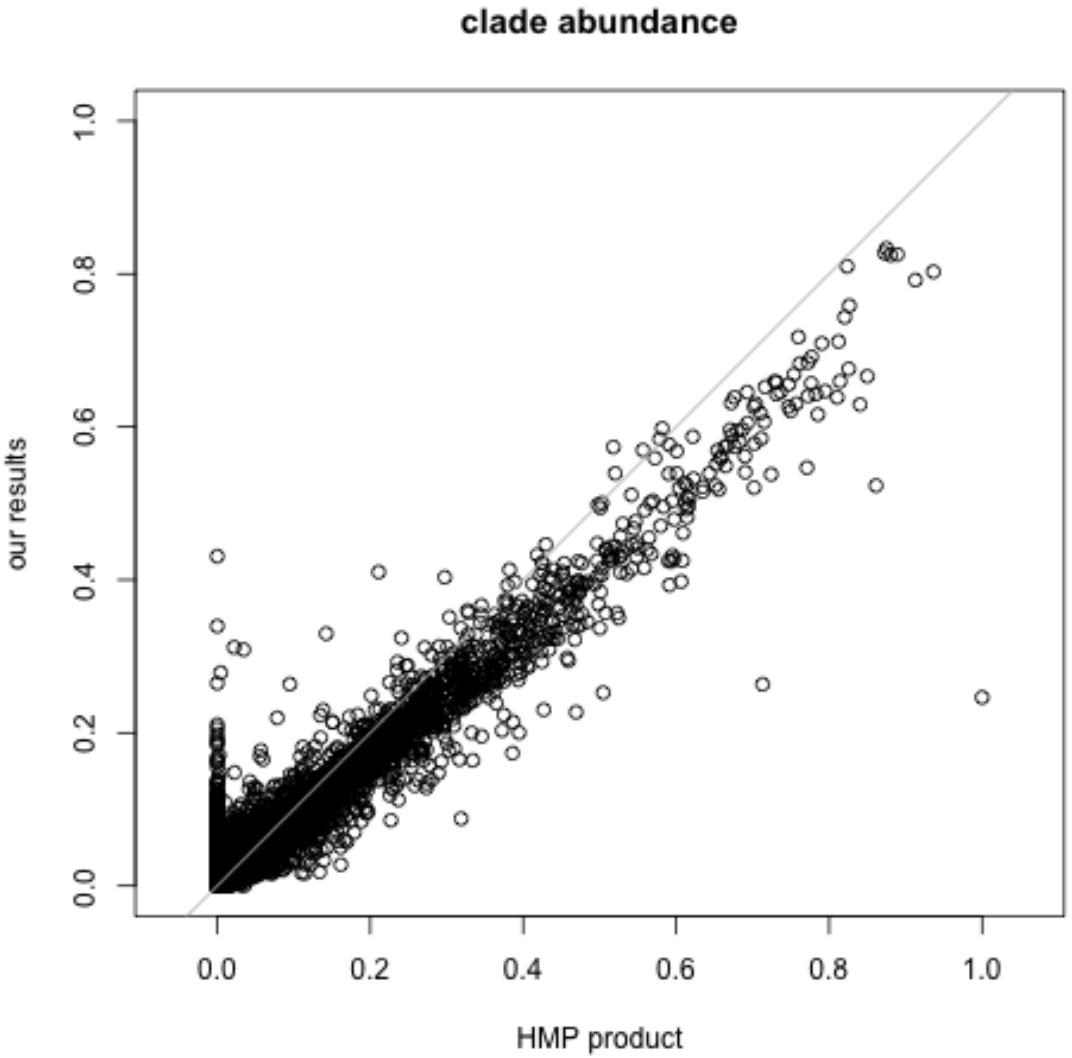
Quality check

Let A_{ij} be the "relative abundance" of species j in sample i, defined as

$$A_{ij} = C(\frac{\sum_{k \in S_j} X_{ik}}{\sum_{k \in S_j} l_k})$$

Taxa abundance A_{ij} is available as a product from the HMP2 pipeline. We can calculate A_{ij} from the marker counts and compare with the HMP2 product.

Quality check



Quality check

- The taxa abundance table provided by HMP2 only have 520 species. It is highly possible that some low abundance species have been filtered out.

Callouts

Why not using the data in the curatedMetagenomicData package?
From the developer:
Given a number n of reads mapping onto a marker (obtained by setting `-t marker_count` in MetaPhlAn), the output in curatedMetagenomicData is equal to

$$v = 1000n/(\text{len\_marker -avg\_read\_len} + 1)$$

Theoretically, one could obtain the number n, by taking the value in curatedMetagenomicData v, and doing:

$$n = (v(\text{len\_marker - avg\_read\_length} + 1))/1000$$

The only issue with this is that the metadata of curatedMetagenomicData reports only the median_read_length, so one could obtain a (good) approximation.
The reason for this (- avg_read_length) is to exclude reads that are outside the limits of the marker.

References I