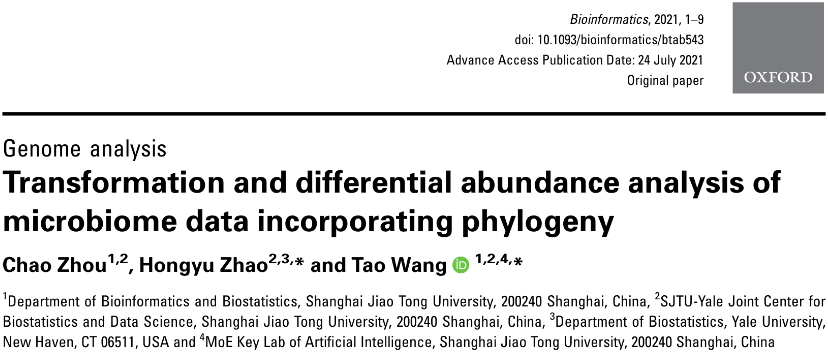
Transformation and differential abundance analysis of microbiome data incorporating phylogeny

Emily Palmer

Oregon State University

palmerem@oregonstate.edu

February 7, 2022



Bioinformatics, 2021, 1–9
doi: 10.1093/bioinformatics/btab543
Advance Access Publication Date: 24 July 2021
Original paper

Genome analysis

Transformation and differential abundance analysis of microbiome data incorporating phylogeny

Chao Zhou[1,2], Hongyu Zhao[2,3,*] and Tao Wang [1,2,4,*]

[1]Department of Bioinformatics and Biostatistics, Shanghai Jiao Tong University, 200240 Shanghai, China, [2]SJTU-Yale Joint Center for Biostatistics and Data Science, Shanghai Jiao Tong University, 200240 Shanghai, China, [3]Department of Biostatistics, Yale University, New Haven, CT 06511, USA and [4]MoE Key Lab of Artificial Intelligence, Shanghai Jiao Tong University, 200240 Shanghai, China

Outline: Overview of paper

- Extend Dirichlet-tree multinomial to zero-inflated Dirichlet tree multinomial to model multivariate microbial counts.
- Use an Empirical Bayes based posterior mean transformation to convert raw counts into non-zero relative abundances that sum to 1.
- Introduce phylogenetically informed DA procedure on transformed data - adaptive analysis of composition of microbiomes for DA testing (adaANCOM)- log-ratios adaptively on the tree for each taxon.

Introduction

- Differential abundance tests need to account for the high dimensional, sparse, compositional, negatively and positively correlated, phylogenetically structured microbiome data.
- Scaling methods to correct for compositional bias (TMM, DESeq, CSS) - assume that most taxa are not differentially abundant, when the count matrix is sparse it can over or underestimate diversity, distort correlations

Review of Dirichlet-Multinomial distribution

- The Dirichlet-multinomial (DM) distribution is commonly used for OTU counts and is a compound distribution comprised of Dirichlet and Multinomial distributions
- Let $\boldsymbol{y} = (y_1, \ldots, y_K)^T$ be the count vector for a sample with K OTUs.
- $y^+ = \sum_{k=1}^K y_k, \boldsymbol{p} = (p_1, \ldots, p_K)^T, \sum_{k=1}^K p_k = 1, \alpha^+ = \sum_{k=1}^K \alpha_k$

- Multinomial pdf:

$$f_M(\boldsymbol{y};\boldsymbol{p}) = \frac{\Gamma(y^+ + 1)}{\prod_{k=1}^K \Gamma(y_k + 1)} \prod_{k=1}^K p_k^{y_k},$$

- Dirichlet pdf:

$$f_D(\boldsymbol{p};\boldsymbol{\alpha}) = \frac{\Gamma(\alpha^+)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^K p_k^{\alpha_k},$$

- Combining both we get the Dirichlet-Multinomial distribution:

$$\begin{aligned} f_{DM}(\boldsymbol{y};\boldsymbol{\alpha}) &= \int f_M(\boldsymbol{y};\boldsymbol{p}) f_D(\boldsymbol{p};\boldsymbol{\alpha}) d\boldsymbol{p} \\ &= \frac{\Gamma(y^+ + 1)\Gamma(\alpha^+)}{\Gamma(y^+ + \alpha^+)} \prod_{k=1}^K \frac{\Gamma(y_k + \alpha_k)}{\Gamma(y_k + 1)\Gamma(\alpha_k)} \end{aligned}$$

Relationship between Gamma, Beta, & Dirichlet distributions

- Let $Z_1, \dots, Z_K$ independent $Z_k \sim \text{Gamma}(\alpha_k, \lambda)$,
- $X_k = \frac{Z_k}{\sum_{j=1}^{K} Z_j}$
- $W_k = \frac{Z_k}{\sum_{j=k}^{K} Z_j}$
- Joint distribution of $\boldsymbol{X} = (X_1, \dots, X_K)^T$ is Dirichlet distributed with parameter $\boldsymbol{\alpha}$
- $W_k \sim \text{Beta}(\alpha_k, \alpha_k^+)$ independent
- $X_k = W_k \prod_{j=1}^{k-1}(1 - W_j)$
- We can rewrite the Dirichlet-Multinomial distribution as

$$f_{DM}(\boldsymbol{y}; \boldsymbol{\alpha}) = \int f_M(\boldsymbol{y}; h(\boldsymbol{w})) \prod_{k=1}^{K-1} f_B(w_k; \alpha_k, \alpha_k^+) d\boldsymbol{w}.$$

- $h(\cdot)$ is the transformation from $\boldsymbol{W}$ to $\boldsymbol{X}$.

Zero Inflated Dirichlet Multinomial

- We can introduce zero inflation by defining the Zero Inflated Dirichlet Multinomial (ZIDM) as:

$$f_{ZIDM}(\boldsymbol{y};\boldsymbol{\pi},\boldsymbol{\alpha}) = \int f_M(\boldsymbol{y}; h(\boldsymbol{w})) \prod_{k=1}^{K-1} f_{ZIB}(w_k;\pi_k,\alpha_k,\alpha_k^+) d\boldsymbol{w}$$

- where

$$f_{ZIB}(w_k;\pi_k,\alpha_k,\alpha_k^+) = \pi_k\delta(0) + (1-\pi_k) f_B(w_k;\alpha_k,\alpha_k^+).$$

  is a zero-inflated binomial distribution

- π_k is the probability of zero-inflation in the kth component $\delta(\cdot)$ is the Dirac delta function

Introducing the Dirichlet Tree Multinomial

- Suppose the relationships between OTUs are encoded in a tree $\mathcal{T}$ which is composed of internal nodes $\mathcal{V}$ and leaf nodes $\mathcal{L}$
- Leaf nodes are OTUs, internal nodes are the node to branch on the phylogenetic tree
- For each $v \in \mathcal{V}$, denote C_v as the set of child nodes of v, $\boldsymbol{y}_v$ the vector of counts correpsonding to C_v, and $y_v^+ = \sum_{u \in C_v} y_u$.

Dirichlet Tree Multinomial & Zero-Inflated Dirichlet Tree Multinomial

- Assume that $\boldsymbol{y}_v$ conditional on y_v^+ are independent across internal nodes, the Dirichlet Tree Multinomial (DTM) distribution is the product of DM distributions that factorize over the tree

$$f_{DTM}(\boldsymbol{y}; \boldsymbol{\alpha}_v, v \in \mathcal{V}) = \prod_{v \in \mathcal{V}} f_{DM}(\boldsymbol{y}_v; y_v^+, \boldsymbol{\alpha}_v)$$

- To extend the DTM to the ZIDTM, we replace the Dirichlet multinomial with Zero-Inflated Dirichlet Multinomial

$$f_{ZIDTM}(\boldsymbol{y}; \boldsymbol{\pi}_v, \boldsymbol{\alpha}_v, v \in \mathcal{V}) = \prod_{v \in \mathcal{V}} f_{ZIDM}(\boldsymbol{y}_v; y_v^+, \boldsymbol{\pi}_v, \boldsymbol{\alpha}_v)$$

Example

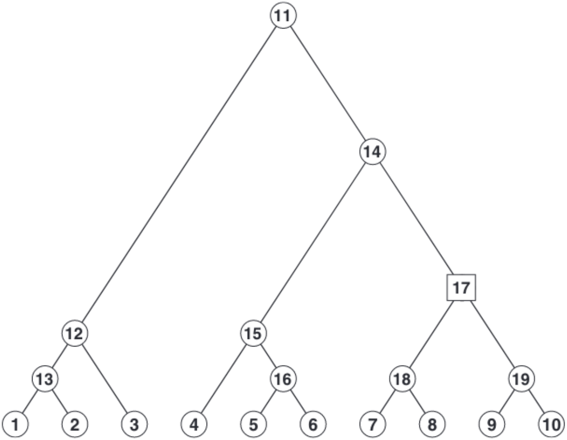


Fig. 1. A binary tree with $K = 10$ leaves. Here $\mathcal{L} = \{1, 2, \ldots, 10\}$, $\mathcal{V} = \{11, 12, \ldots, 19\}$. For illustration, $\mathcal{C}_{17} = \{18, 19\}, \boldsymbol{y}_{17} = (y_{18}, y_{19})^T$ and $y_{17}^+ = y_{18} + y_{19}$. Given y_{17}^+, $\boldsymbol{y}_{17}$ has a DM or ZIDM distribution. The factorization over internal nodes means that these conditional distributions are independent

Maximum likelihood estimation for ZIDTM

- Use EM-algorithm to estimate unknown parameters: $\boldsymbol{\theta} = \{\boldsymbol{\alpha}_v, \boldsymbol{\pi}_v, v \in \mathcal{V}\}$
- Assume $C_v = \{1, \ldots, K_v\}$, where $K_v = |C_v|$
- Consider n observations $\boldsymbol{y}^1, \ldots, \boldsymbol{y}^n$, then the log-likelihood function (without constant terms) is:

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{n} \sum_{v \in \mathcal{V}} \sum_{k=1}^{K_v - 1} \{l_1(\delta_{vk}^i, \pi_{vk}) + l_2(\delta_{vk}^i, w_{vk}^i, \alpha_{vk}, \alpha_{vk}^+)\}$$

with

$$l_1(\delta_{vk}^i, \pi_{vk}) = \delta_{vk}^i \log \pi_{vk} + (1 - \delta_{vk}^i) \log(1 - \pi_{vk})$$

$$l_2(\delta_{vk}^i, w_{vk}^i, \alpha_{vk}, \alpha_{vk}^+) = (1 - \delta_{vk}^i)\{-\log \mathcal{B}(\alpha_{vk}, \alpha_{vk}^+) + (\alpha_{vk} - 1) \log w_{vk}^i + (\alpha_{vk}^+ - 1) \log(1 - w_{vk}^i)\}$$

- δ_{vk}^i is the indicator of zero-inflation $\mathcal{B}()$ is the beta function

E-step

- Compute expectation of $l(\boldsymbol{\theta})$ with respect to posterior distribution of $(\delta^i_{vk}, w^i_{vk})|\boldsymbol{y}^i_v$, indexed by current value of $\boldsymbol{\theta}$, which gives the Q-function:

$$Q(\boldsymbol{\theta}) = \sum_{i=1}^{n} \sum_{v \in \mathcal{V}} \sum_{k=1}^{K_v-1} E\{l_1(\delta^i_{vk}, \pi_{vk}) + l_2(\delta^i_{vk}, w^i_{vk}, \alpha_{vk}, \alpha^+_{vk})\}$$

E-step: Definitions

- $\delta_{vk}^{i*} = E(\delta_{vk}^{i} | \boldsymbol{y}_{v}^{i})$
- $R_{vk}^{i*} = E(\log w_{vk}^{i} | \boldsymbol{y}_{v}^{i}, \delta_{vk}^{i} = 0)$
- $S_{vk}^{i*} = E\{\log(1 - w_{vk}^{i}) | \boldsymbol{y}_{v}^{i}, \delta_{vk}^{i} = 0\}$
- Then,

$$\delta_{vk}^{i*} = \begin{cases} 0, & y_{vk}^{i} > 0, \\ \dfrac{\pi_{vk}}{\pi_{vk} + (1 - \pi_{vk}) \dfrac{\mathcal{B}(\alpha_{vk}^{i*}, \alpha_{vk}^{i*+})}{\mathcal{B}(\alpha_{vk}, \alpha_{vk}^{+})}}, & y_{vk}^{i} = 0, \end{cases}$$

$$S_{vk}^{i*} = \psi(\alpha_{vk}^{i*+}) - \psi(\alpha_{vk}^{i*} + \alpha_{vk}^{i*+}) \qquad R_{vk}^{i*} = \psi(\alpha_{vk}^{i*}) - \psi(\alpha_{vk}^{i*+})$$

- $\alpha_{vk}^{i*} = \alpha_{vk} + y_{vk}^{i}, \quad \alpha_{vk}^{i*+} = \sum_{j=k+1}^{K_v} \alpha_{vk}^{i*}, \quad \psi(\cdot)$ is the digamma function.

E-step: Q-function

- We can rewrite $Q(\boldsymbol{\theta})$ as:

$$Q(\boldsymbol{\theta}) = \sum_{i=1}^{n} \sum_{v \in \mathcal{V}} \sum_{k=1}^{K_v - 1} \{Q_1(\pi_{vk}, \delta^{i}_{vk}) + Q_2(\alpha_{vk}, \alpha^{+}_{vk}, R^{i*}_{vk}, S^{i*}_{vk})\}$$

- where

$$Q_1(\pi_{vk}, \delta^{i}_{vk}) = \delta^{i*}_{vk} \log(\pi_{vk}) + (1 - \delta^{i*}_{vk}) \log(1 - \pi_{vk})$$

and

$$Q_2(\alpha_{vk}, \alpha^{+}_{vk}, R^{i*}_{vk}, S^{i*}_{vk}) = (1 - \delta^{i*}_{vk})\{-\log \mathcal{B}(\alpha_{vk}, \alpha^{+}_{vk})$$

$$+(\alpha_{vk} - 1)R^{i*}_{vk} + (\alpha^{+}_{vk} - 1)S^{i*}_{vk}\}$$

M-step

- Maximize $Q(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$
- The parameters in Q_1 and Q_2 can be optimized separately
- Model depends on ordering of OTUs at each internal node (matching problem)
- Fit separate ZIDM model for each possible ordering of $|C_v|$ taxa, and select the best fitted model.
- Computational cost is $O(\sum_{v \in \mathcal{V}} |C_v|!)$ compared to $O(|\mathcal{L}|!)$ for GDM or ZIDM models

Binary trees

- From now on, consider binary trees
- i.e. $|C_v| = 2$ for all $v \in \mathcal{V}$
- Dirichlet, multinomial, dirichlet multinomial, and zero inflated dirichlet multinomial are reduced to beta, binomial, beta-binomial and zero-inflated beta-binomial.
- The ZIDTM is then a product of ZIBBs

Posterior mean transformation

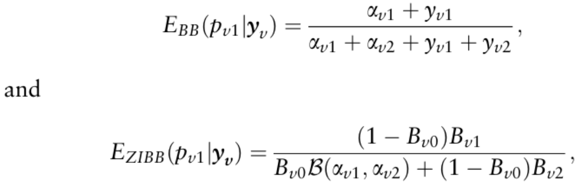
- Estimate underlying proportions from Bayesian perspective (since Dirichlet is conjugate to multinomial)
- Consider the posterior mean for the Dirichlet Multinomial:

$$E_{DM}(p_k|\mathbf{y}) = \frac{\alpha_k + y_k}{\sum_{j=1}^{K}(y_j + \alpha_j)}$$

- We estimate unknown parameters by maximizing data likelihood
- Use the estimated parameters as ”pseudo data” and combine with real observed data.
- This produces non-zero proportions for zero counts
- Alternative to using a pseudo-count for sample proportions
- Becomes difficult in presence of zero-inflation

Solving posterior mean transformation

- When we have a binary tree, there is an explicit closed-form solution to the zero-inflation
- At each internal node $v \in \mathcal{V}$,



$$E_{BB}(p_{v1}|\boldsymbol{y}_v) = \frac{\alpha_{v1} + y_{v1}}{\alpha_{v1} + \alpha_{v2} + y_{v1} + y_{v2}},$$

and

$$E_{ZIBB}(p_{v1}|\boldsymbol{y}_v) = \frac{(1 - B_{v0})B_{v1}}{B_{v0}\mathcal{B}(\alpha_{v1}, \alpha_{v2}) + (1 - B_{v0})B_{v2}},$$

  - $B_{v0} = \pi I(y_{v1} = 0)$
  - $B_1 = \mathcal{B}(1 + \alpha_{v1} + y_{v1}, \alpha_{v2} + y_{v2})$
  - $B_2 = \mathcal{B}(\alpha_{v1} + y_{v1}, \alpha_{v2} + y_{v2})$

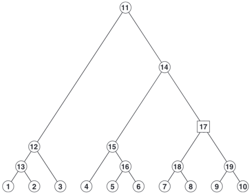
Likelihood-ratio test & count transformation

- Having a correct model specification at internal nodes effects quality of posterior estimates.
- Perform a two-stage likelihood-ratio test
- First, assume count data at $v \in \mathcal{V}$ are not zero-inflated, fit beta-binomial model and test for over-dispersion.
- Counts at nodes without over-dispersion are transformed into sample proportions after adding constant of .5, equivalent to using E_{BB} with $\alpha_{v1} = \alpha_{v2} = 0.5$
- Second, Nodes with over-dispersion are refit using a ZIBB model, and tested again for zero-inflation
- Counts are then transformed using $E_{ZIBB}(p_{v1}|\boldsymbol{y}_v)$.
- If there is no zero-inflation, counts are transformed using $E_{BB}(p_{v1}, \boldsymbol{y}_v)$
- $\hat{\boldsymbol{p}}_v = (\hat{p}_{v1}, 1 - \hat{p}_{v1})^T$ is the posterior estimate

Path-level definitions - Phylogeny-aware normalization

- To ensure that the normalization is 'phylogeny-aware', we consider the path level instead of focussing on individual internal nodes.
- Define $\mathcal{A}_u$ as the ancestor node set of u for each node $u \in \mathcal{L} \cup \mathcal{V}$ that contains all internal nodes in the path from the root node to u
- $\mathcal{L}_u$ is the set of leaves in the same path

- $\mathcal{A}_1 = \mathcal{A}_2 = \{11, 12, 13\}$, $\mathcal{A}_3 = \{11, 12\}$, $\mathcal{L}_{11} = \mathcal{L}, \mathcal{L}_{12} = \{1, 2, 3\}, \mathcal{L}_{13} = \{1, 2\}$



- Define $q_u = \prod_{v \in \mathcal{A}_u} \hat{p}_v$
- $\mathcal{U}$ is the set of nodes such that $\cup_{u \in U} \mathcal{L}_u = \mathcal{L}$, $\mathcal{L}_u \cap \mathcal{L}_{u'} = \emptyset$ when $u \neq u'$.
- Then, $\sum_{u \in \mathcal{U}} q_u = 1$ and $\sum_{l \in \mathcal{L}} q_l = 1$

adaANCOM

- Detect differentially abundant OTUs at ecosystem level
- Similar to ANCOM
- Difference is how we use the phylogenetic tree to incorporate this information
- Tests differences between two groups
- Speed up ANCOM by using phylogenetic tree to construct adaptive log-ratios

Hypothesis tests - ANCOM

- Test the hypothesis

  $$H_{0l} : \log \mu_l^1 = \log \mu_l^2$$

  for $l \in \mathcal{L} = \{1, \ldots, K\}$, and where μ_l^g is the mean absolute abundance of the lth OTU from gth group, and $g = 1, 2$
- In ANCOM, we test $K - 1$ hypothesis using each OTU as the denominator in the additive log ratio

  $$H_{0lm} : \log(\mu_l^1 / \mu_m^1) = \log(\mu_l^2 / \mu_m^2)$$

  for all $m \neq l$
- ANCOM counts the number of rejections across the $K - 1$ hypothesis tests, and compares to the empirical distribution of counts to determine if the lth OTU is differentially abundant.
- If K is high, ANCOM suffers from a high FDR. Additionaly the computational burden is high

adaANCOM log-ratios & hypothesis tests

- Assume that abundance difference on the log scale at an internal node is passed down to child nodes
- adaANCOM is comprised of a two-step process
  1. Test the internal node-level hypothesis

     $$H_{0v} : \log(\mu_{v1}^1/\mu_{v2}^1) = \log(\mu_{v1}^2/\mu_{v2}^2)$$

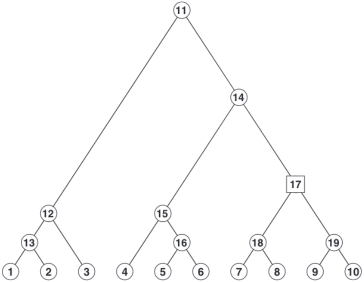
     for $v \in \mathcal{V}$, with μ_{v1}^g and μ_{v2}^g as mean absolute abundances of children of v from the gth group.
     For a given α,$\mathcal{D}_\mathcal{V}$ is the set of internal nodes which the hypotheses are rejected
  2. For each leaf node $l \in \mathcal{L}$, calculate the log-ratio
     Define ref_l as the sibling node of l if $\mathcal{A}_l \cap \mathcal{D}_\mathcal{V} = \emptyset$ and the child node of v not in $\mathcal{A}_l$ otherwise, when v is the node in $\mathcal{A}_l \cap \mathcal{D}_\mathcal{V}$ closest to the root node.
     The null hypothesis is:

     $$H_{0l}^{ada} : \log(\mu_l^1/\mu_{ref_l}^1) = \log(\mu_l^2/\mu_{ref_l}^2)$$

- adaANCOM rejects the null hypothesis H_{0l} if H_{0l}^{ada} is rejected

adaANCOM example

- Assume $\mathcal{D}_{\mathcal{V}} = 17$.
- $ref_1 = 2, ref_2 = 1$
  $ref_3 = 13, ref_4 = 16,$
  $ref_5 = 6, ref_6 = 5,$
  $ref_7 = ref_8 = 19,$
  $ref_9 = ref_{10} = 18$



Accounting for log-ratios of zero y_u counts

- H_{0v} and H_{0l}^{ada} are based on the log-ratios of q_us for $u \in \mathcal{L} \cup \mathcal{V}$, but when the original y_u counts have many zero counts, log-ratios give abnormal results, and test statistics are sensitive to these abnormal values
- For internal node v, define ϕ_v as the maximum of

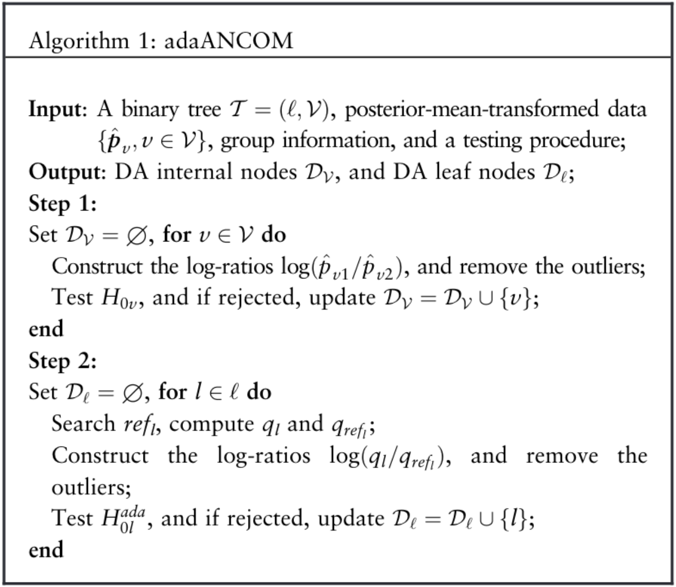
$$|\log(\hat{p}_{v1}/\hat{p}_{v2})|$$

  that have $y_{v1} > 0$ and $y_{v2} > 0$.
- Data with $y_{v1} = 0$ or $y_{v2} = 2$ are removed if $|\log(\hat{p}_{v1}/\hat{p}_{v2})| > \phi_v$

Comparisons to ANCOM

- Similar to ANCOM, adaANCOM uses relative abundance data, constructs log-ratios, and performs t-tests or Wilcoxon-rank-sum tests.
- The advantage of adaANCOM is the computation time, as the number of tests is reduced by a factor of $|\mathcal{L}|$.
- adaANCOM also has an advantage of interpretability since we are guided by the tree, so we result in both DA leaves and nodes.
- Additionally, adaANCOM controls FDR better

adaANCOM algorithm



Algorithm 1: adaANCOM

```
Input: A binary tree $\mathcal{T} = (\ell, \mathcal{V})$, posterior-mean-transformed data
       $\{\hat{p}_v, v \in \mathcal{V}\}$, group information, and a testing procedure;
Output: DA internal nodes $\mathcal{D}_{\mathcal{V}}$, and DA leaf nodes $\mathcal{D}_{\ell}$;
Step 1:
Set $\mathcal{D}_{\mathcal{V}} = \varnothing$, for $v \in \mathcal{V}$ do
    Construct the log-ratios $\log(\hat{p}_{v1}/\hat{p}_{v2})$, and remove the outliers;
    Test $H_{0v}$, and if rejected, update $\mathcal{D}_{\mathcal{V}} = \mathcal{D}_{\mathcal{V}} \cup \{v\}$;
end
Step 2:
Set $\mathcal{D}_{\ell} = \varnothing$, for $l \in \ell$ do
    Search $ref_l$, compute $q_l$ and $q_{ref_l}$;
    Construct the log-ratios $\log(q_l/q_{ref_l})$, and remove the outliers;
    Test $H_{0l}^{ada}$, and if rejected, update $\mathcal{D}_{\ell} = \mathcal{D}_{\ell} \cup \{l\}$;
end
```