treeclimbR pinpoints the data-dependent resolution of hierarchical hypotheses

Gian Marco Visani

Huang, R., Soneson, C., Germain, PL. *et al. treeclimbR* pinpoints the data-dependent resolution of hierarchical hypotheses. *Genome Biol* 22, 157 (2021). https://doi.org/10.1186/s13059-021-02368-1

Outline

1. Quick overview of statistical tests and multiple hypotheses problem
2. Justification for treeclimbR
3. treeclimbR algorithm
4. Evaluation of treeclimbR
5. Discussion

Statistical tests in biology

In biology, we often perform statistical tests to infer if there are any differences between two groups (e.g. control group vs. treatment group). Tests are run on values of samples associated with each group.

These tests usually make a null hypothesis, and return whether such null hypothesis is accepted or rejected, and come with a p-value associated with them.

p-value = probability of obtaining values at least as extreme as those given assuming the null hypothesis is correct.

In other words, if we reject the null hypothesis, the p-value is the probability that we are wrong, i.e. the probability of having a false positive (or false discovery).

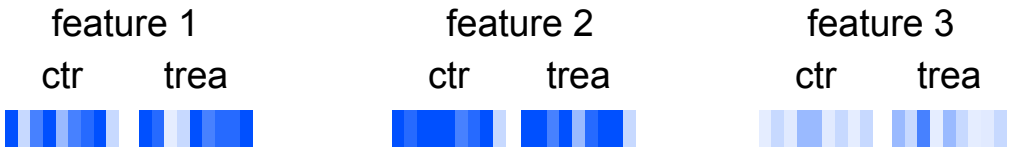
Statistical Tests in biology

Differential Abundance (DA) analysis

- Two sets of samples (control and treatment); one value for each sample
- Null hypothesis: no difference in abundance between control and treatment samples
- Returns “+” (treatment more abundant) or “-” (treatment less abundant) and a p-value

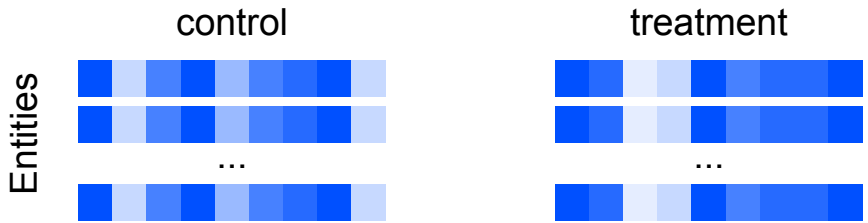
Differential State (DS) analysis

- Same as DA, but multiple features associated with each sample (a “state”)



Multiple Hypotheses Tests problem

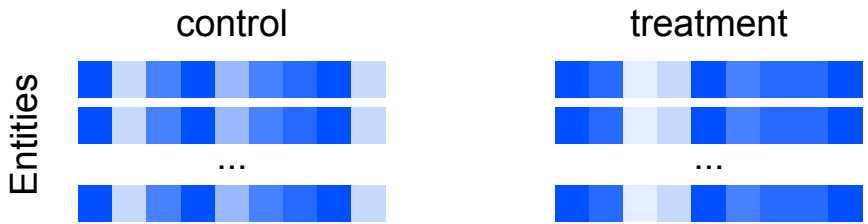
Problems arise when we conduct multiple such tests at the same time



If we reject 1000 null hypotheses each with a p-value of 0.05, we would expect 50 such measurements to be false positives by chance alone, which is not great

Multiple Hypotheses Tests problem

Problems arise when we conduct multiple such tests at the same time



If we reject 1000 null hypotheses each with a p-value of 0.05, we would expect 50 such measurements to be false positives by chance alone, which is not great

False Discovery Rate (FDR): expected number of false discoveries (incorrect rejections of the null) over all discoveries (all rejections of the null)

Benjamin-Hochberg (BH) [1]**:** standard method to control the FDR. Reject null hypotheses keeping FDR below a certain threshold (nominal FDR)

[1] Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple hypothesis testing. J R Stat Soc B 57:289–300

What about the True Positive Rate?

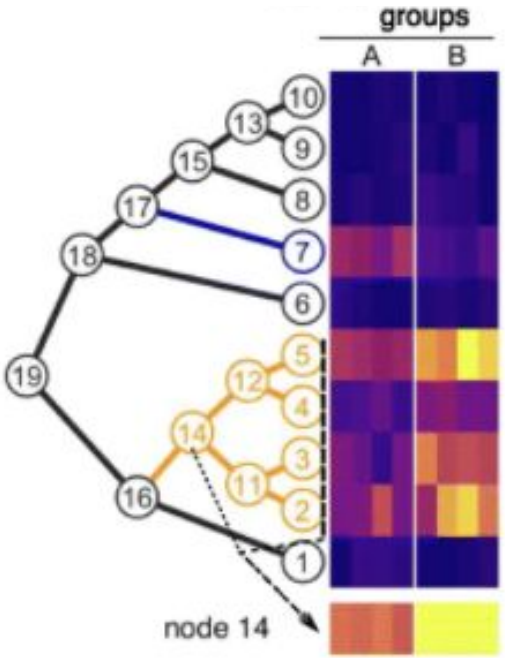
False Negatives: when the null hypothesis is truly false, there might not be enough signal for a statistical test to be deemed significant

Example: in DA analysis, two entities may have low abundance or low fold change or not enough samples, so the differential analysis returns the right sign, but with too high a p-value

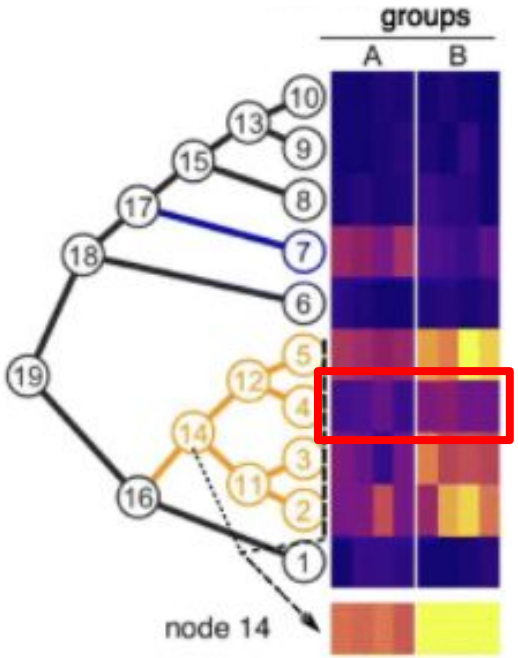
Causes?

- Experimental (not enough data was collected)
- Intrinsic (signal is simply very low for that entity)

Solution - use hierarchy as side information

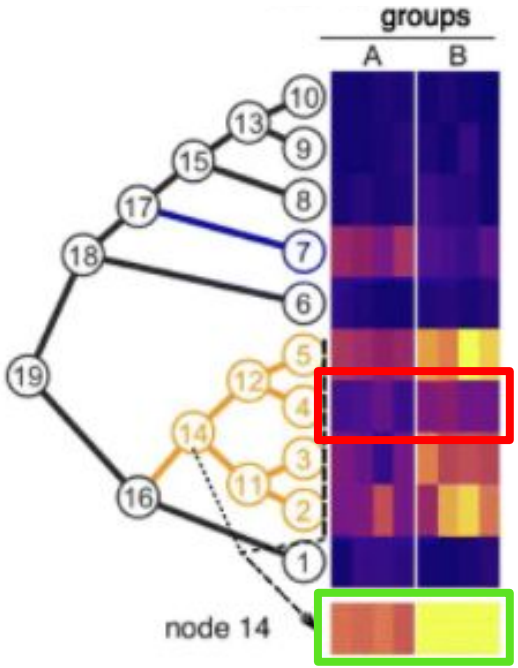


Solution - use hierarchy as side information



True but low signal, false negative

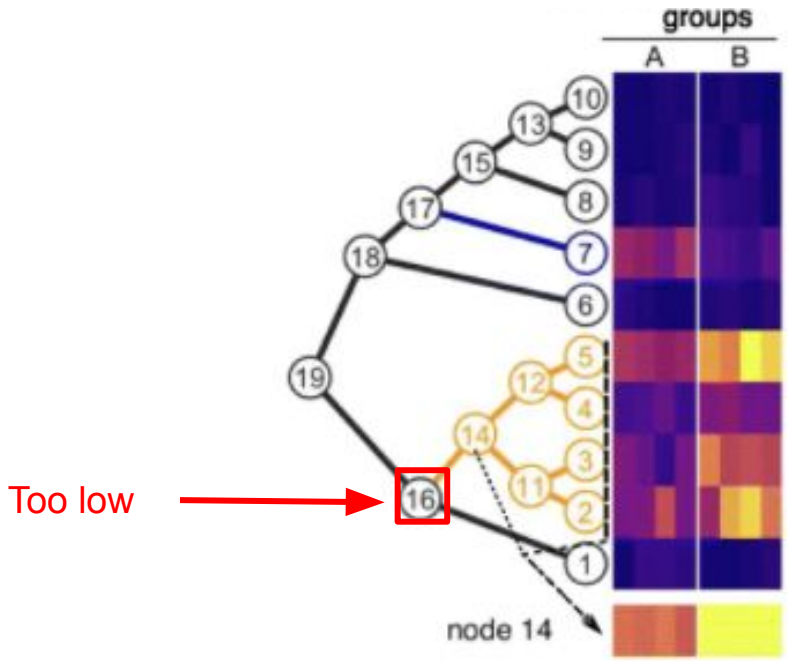
Solution - use hierarchy as side information



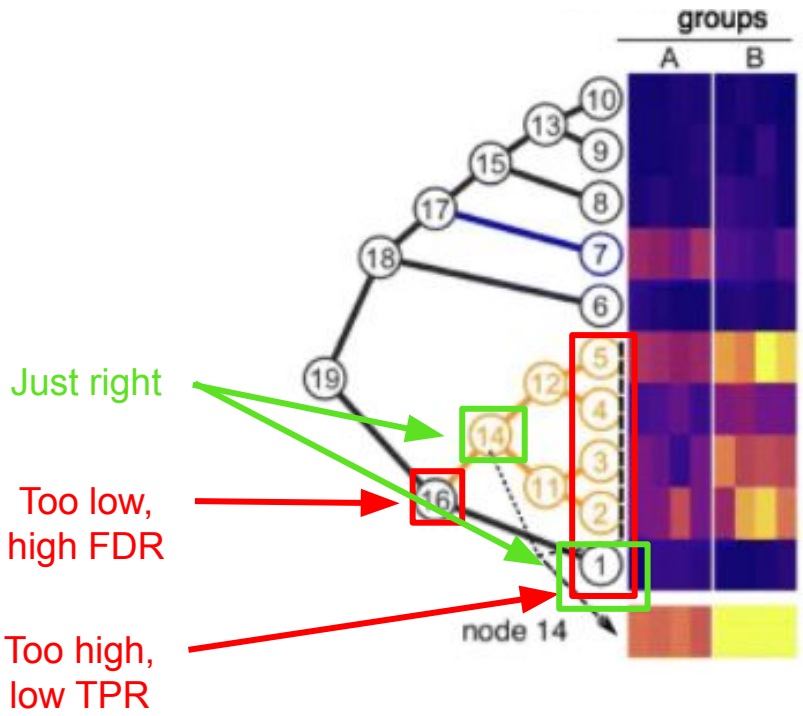
True but low signal, false negative

True and strong signal, true positive

Control the FDR



Finding the right resolution - FDR and TPR tradeoff



Previous work

1. HFDR
2. StructFDR
3. MiLineage
4. Phylofactor
5. LEfSE
6. TASSO
7. rare
8. Citrus
9. diffcyt

...

Each has at least one of the following problems:

1. Only for specific kind of data (e.g. microbiome)
2. Can’t handle DS case
3. Predicts nested nodes
    a. messy interpretation, but not as big of a problem as the authors make it seem

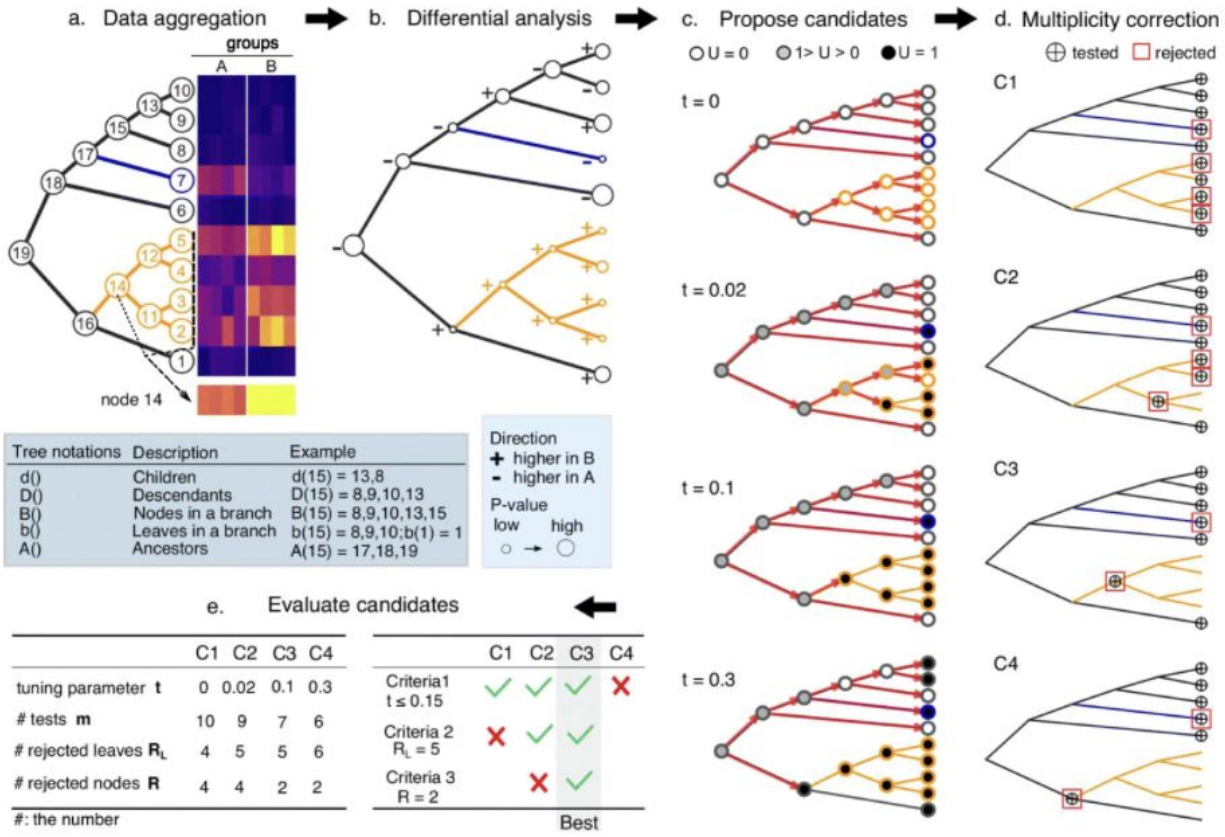
treeclimbR

Five steps:

1. Data aggregation
2. Differential analysis
3. Propose candidates
4. Multiplicity correction
5. Evaluate candidates

Output:

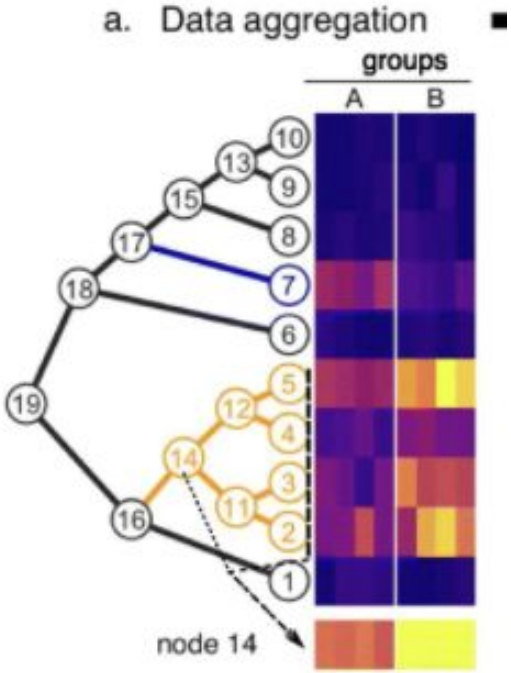
Set of nodes for which null is rejected



Step 1: Data aggregation

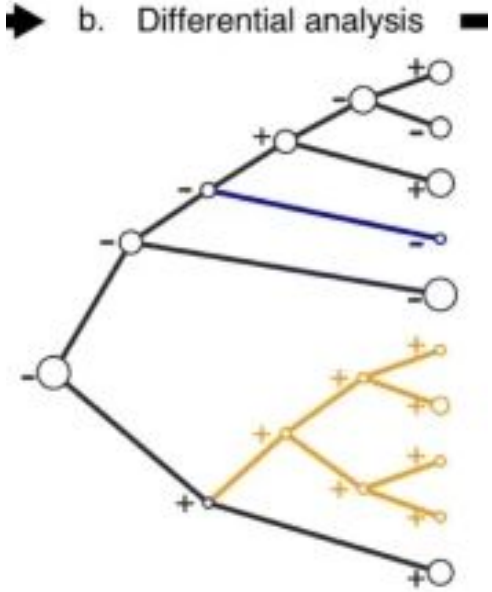
Use mean, median or sum of data “depending on the context”. Authors way too vague.

For DA analysis, sum seems like a good choice.



Step 2: Differential analysis

- Perform DA analysis on each node
- Assigns a **sign/direction** and a **p-value** to each node



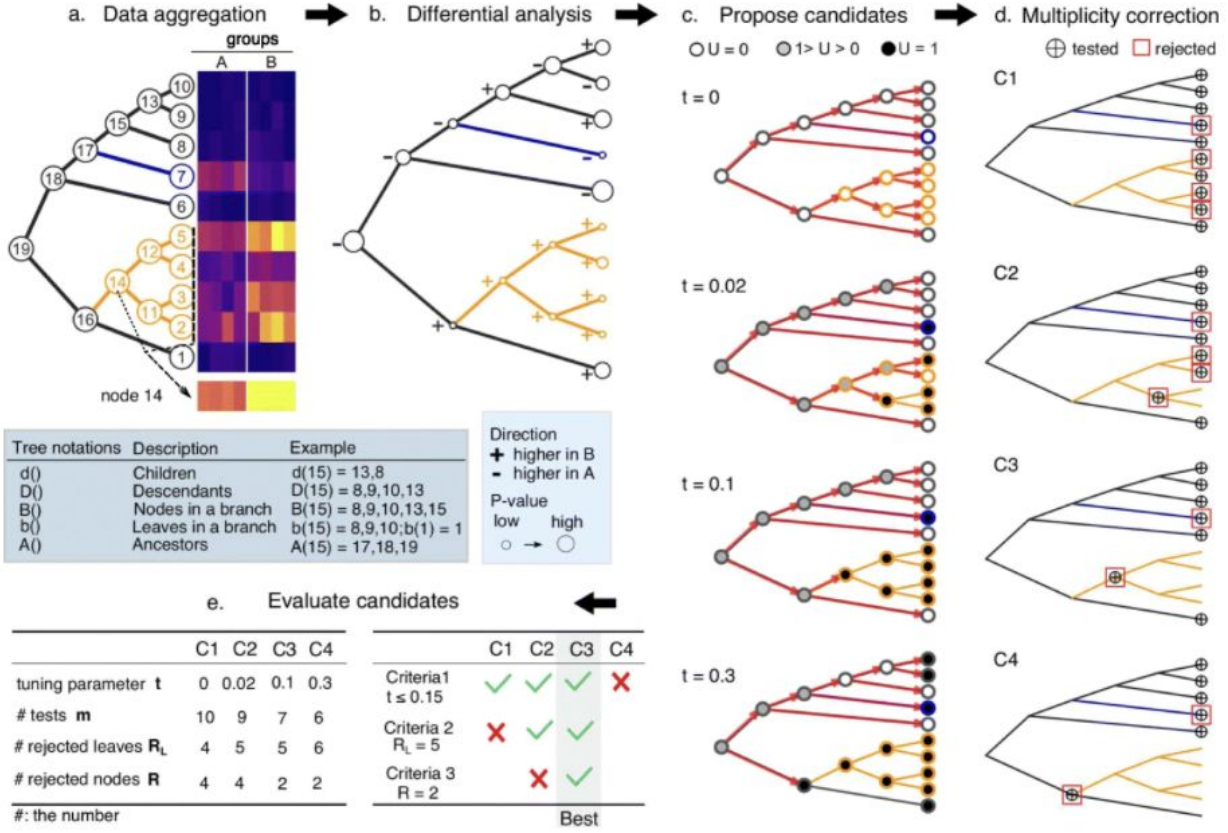
treeclimbR

Five steps:

1. Data aggregation
2. Differential analysis
3. **Propose candidates**
4. **Multiplicity correction**
5. **Evaluate candidates**

Output:

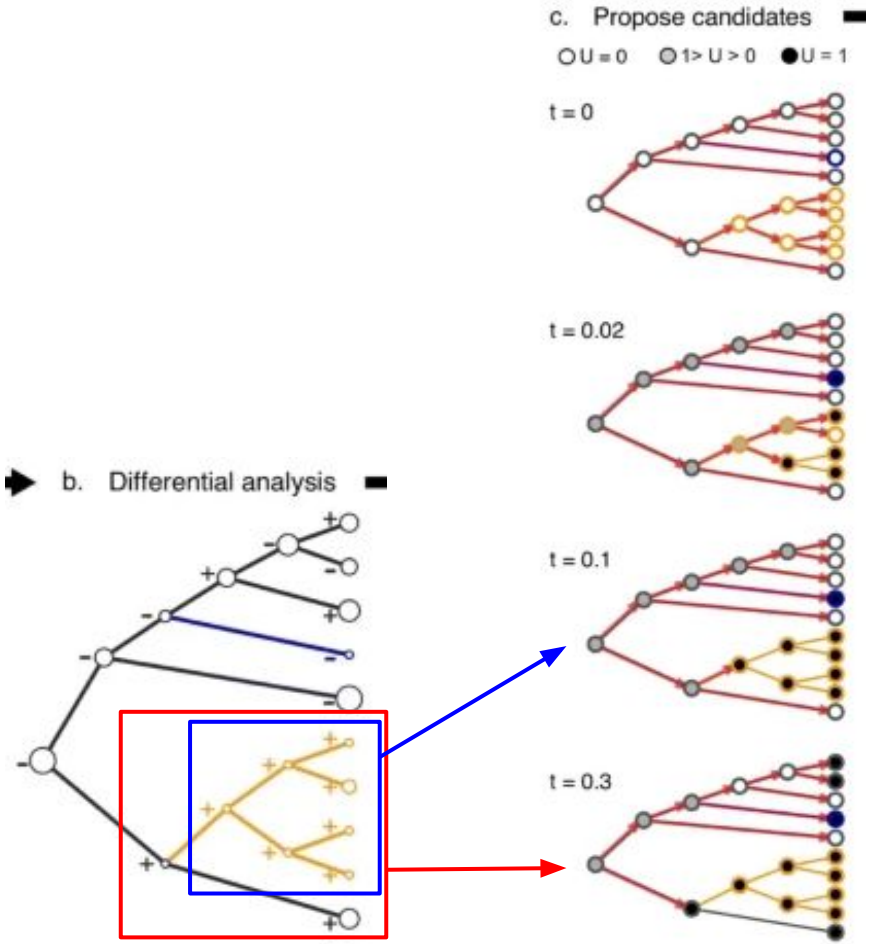
Set of nodes for which null is rejected



Step 3: Propose candidates

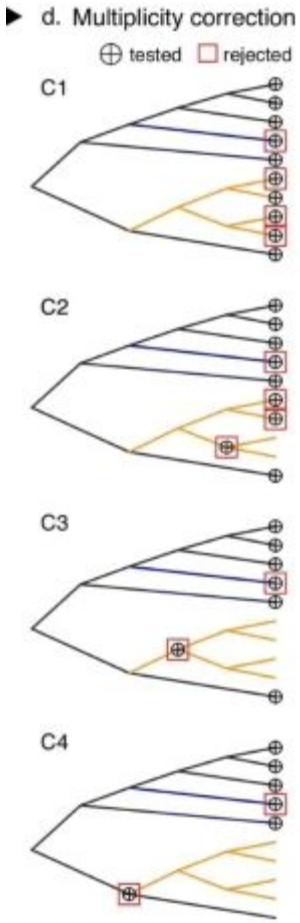
Propose candidate by varying tuning parameter $t \in [0, 1]$

Core idea: stop at a low resolution node **if and only if** all its descendants agree on a direction and each with p-value less than *t*



Step 4: Multiplicity correction

Simply perform BH on each candidate



Step 5: Candidate evaluation

Criterion 1: upper bound on *t* to control the FDR

Candidate is kept if $t < 2\alpha(\frac{l_t}{s_t} - 1)$, where

s_t = number of nodes where null was rejected for this candidate

l_t = number of leaves descending from these rejected nodes

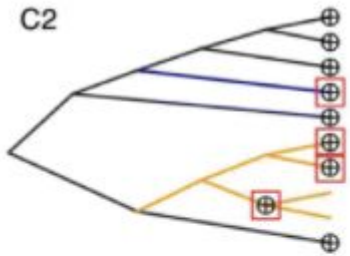
α = nominal FDR

If *t* is kept below this value, then the expected FDR at the leaf level is below α

Step 5: Candidate evaluation

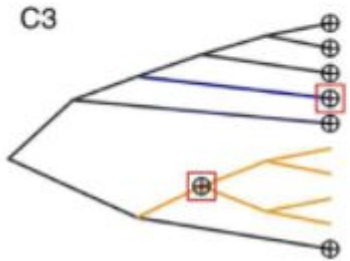
Criterion 2: select candidate that has rejected the null on the highest number of leaves

- Now that FDR is under control, let's increase TPR by rejecting as many null hypotheses as possible



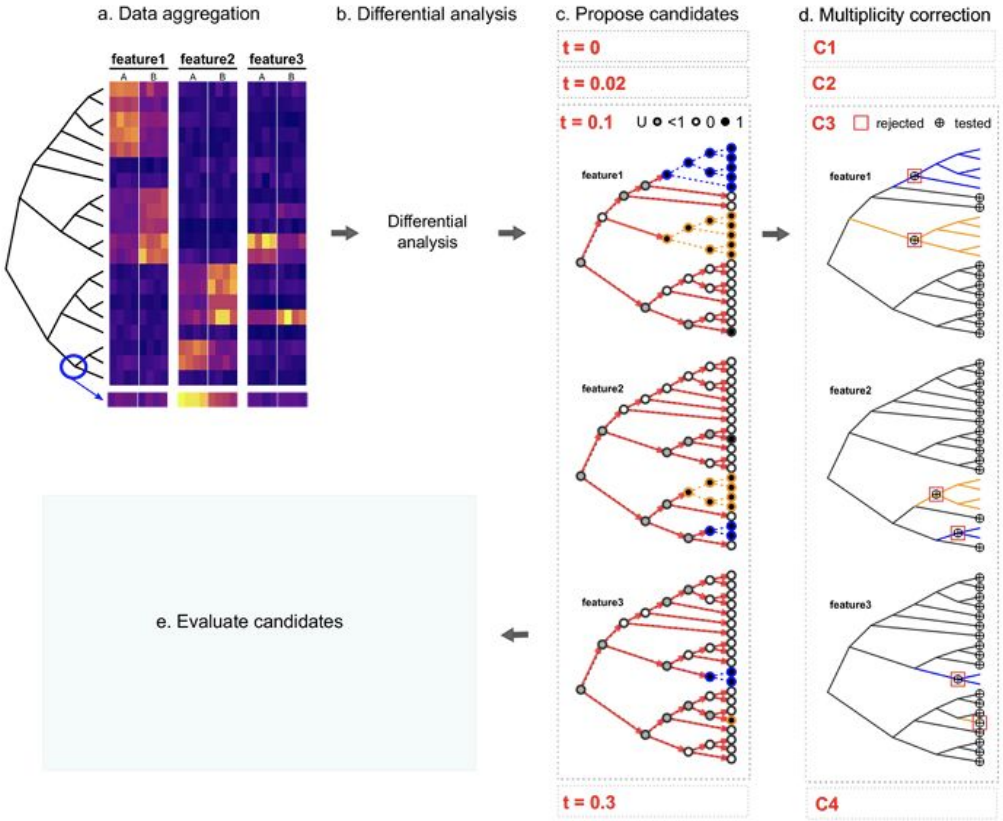
Criterion 3: select the candidate with the least number of rejected nodes

- Essentially, select candidate with lowest resolution among candidates → better interpretability



Questions?

DS analysis case



1. Run steps 2, 3 and 4 independently for each feature
2. Combine the feature-specific candidates via union of nodes:

$$\mathbf{C}(t) = \bigcup_{g \in \mathbf{G}} C_g(t)$$

3. Run candidate evaluation

Node importance?

Evaluation on synthetic data

1. Parametric synthetic microbial datasets
    a. Informative tree
2. Non-parametric synthetic microbial datasets
    a. Uninformative tree
    b. Correlation tree
3. AML-sim
4. BCR-XL-sim

Parametric synthetic microbial datasets

Entities: microbial taxa (OTUs)
Data: OTU counts
Tree: phylogenetic tree

OTU counts are sampled from multinomial distribution with control parameters inferred from real data, and treatment parameters computed to generate three specific scenarios, each with two signal branches

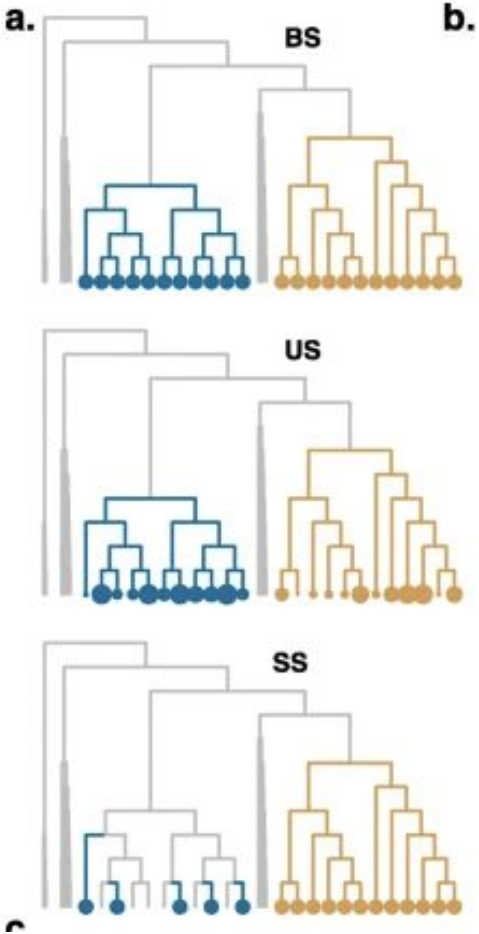
9 datasets: 3 scenarios and 3 sample sizes

BS $\begin{cases} \hat{\pi}_k^T = \hat{\pi}_k^C; & k \notin A, B \\ \hat{\pi}_k^T = r\hat{\pi}_k^C; & k \in A \\ \hat{\pi}_k^T = \frac{1}{r}\hat{\pi}_k^C; & k \in B \end{cases}$

k is an OTU

A, B are the two signal branches

r is the fold change

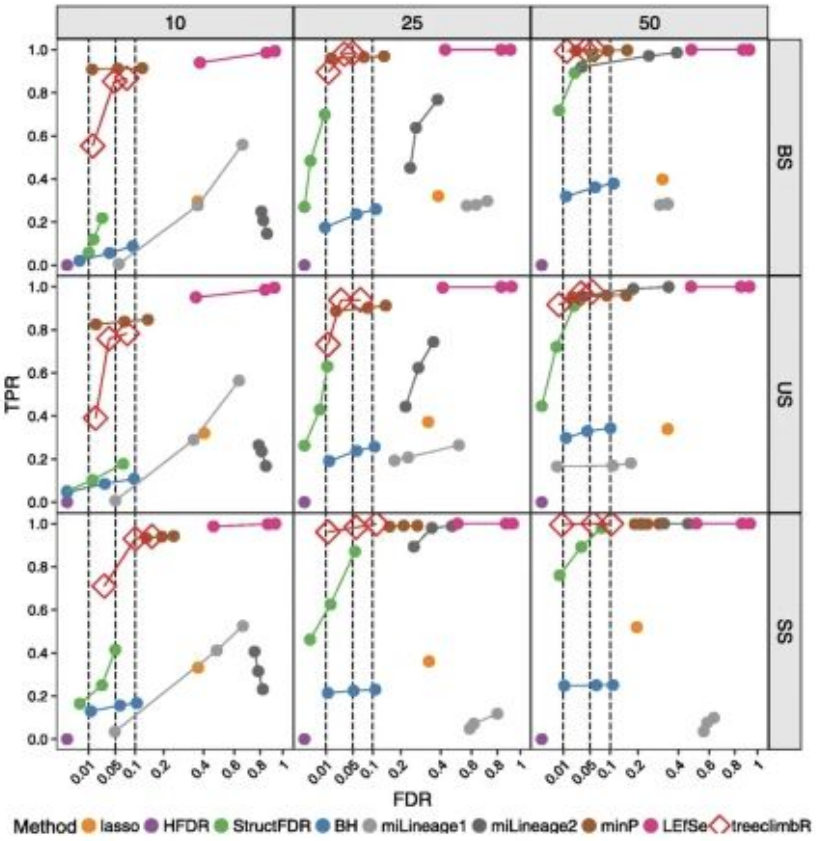


Parametric synthetic microbial datasets

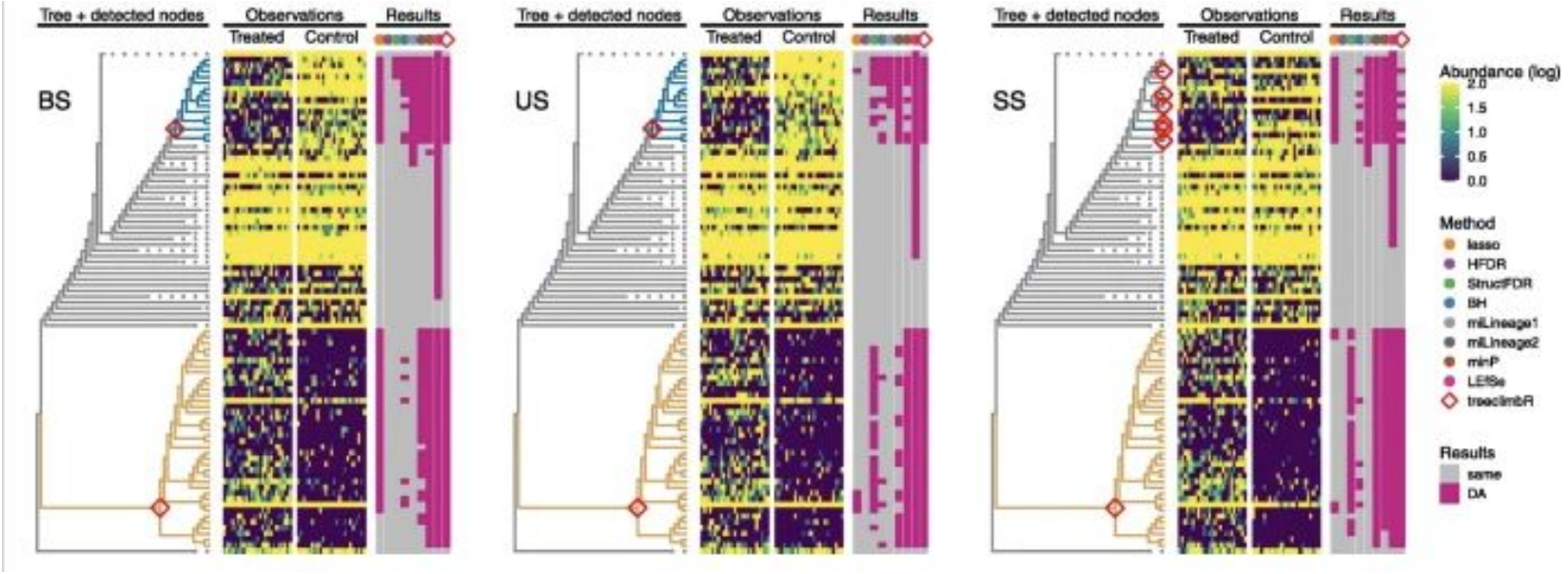
Each method has three points, one for each nominal FDR

Average TPR and true FDR over 100 simulations

treeclimbR has great FDR control, similar to BH, and with high TPR



Parametric synthetic microbial datasets



Non-parametric synthetic microbial datasets

Same control data, two kinds of datasets:

1. Uninformative tree
2. Correlation tree

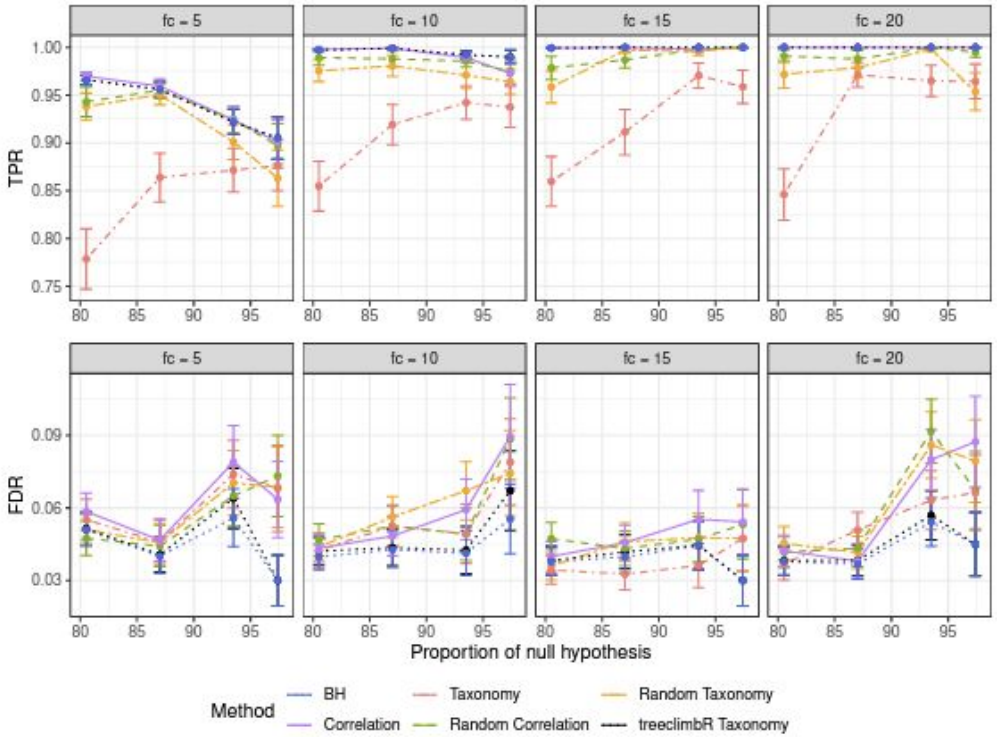
Uninformative tree

Simulate treatment data by selecting OTUs at random to multiply by a fold change

Use original phylogenetic tree

Compared treeclimbR with StructFDR and BH

Main result: if the tree is uninformative, the performance of treeclimbR is analogous to that of BH (i.e. not using the tree), whereas the performance of StructFDR deteriorates



Correlation tree

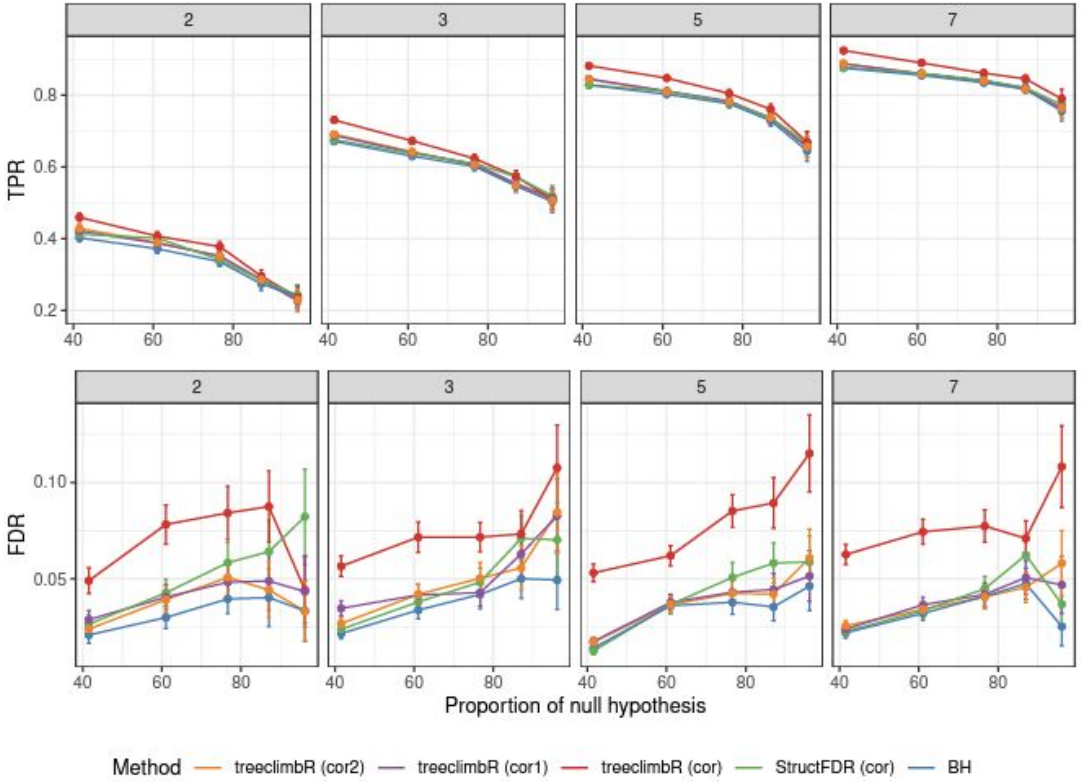
Simulate treatment data by selecting OTUs at random to multiply by a fold change

Construct tree using similarity matrix computed on OTU counts

- **Cor1:** tree built using control data
- **Cor2:** tree built using treatment data
- **Cor:** tree built using both control and treatment data

This tree tends to put in the same branch entities that are differentially abundant in the same direction even if by chance.

- Overestimates t and leads to poor FDR control



AML-sim

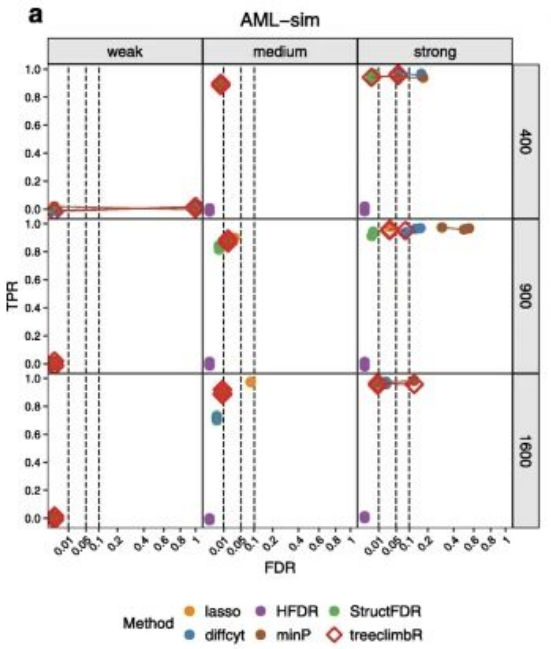
Dataset that simulates the phenotype of minimal residual disease of AML patients

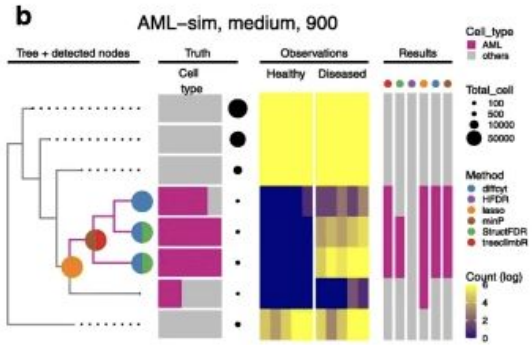
Entities: cell clusters; clustered CyTOF profiles according to lineage markers
Data: cell counts
Tree: generated by hierarchical clustering on a similarity matrix computed on median expression of lineage markers between groups

DA analysis

9 datasets: 3 scenarios and 3 sample sizes

Groups:
- Control
- Diseased (AML)





BCR-XL-sim

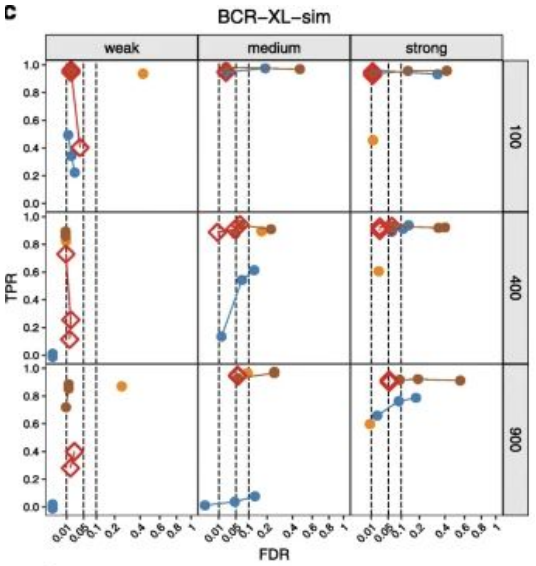
Entities: groups of PBMCs
Data: expression of several protein markers
Tree: generated by hierarchical clustering on a similarity matrix computed on median expression of lineage markers between groups

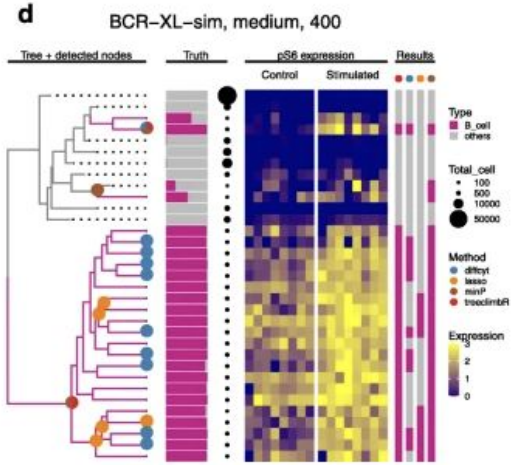
DS analysis

9 datasets: 3 scenarios and 3 sample sizes

Groups:

- Healthy
- Stimulated with B cell receptor cross-linker





Analyses on three real datasets

Goal: show how treeclimbR is able to detect meaningful differential abundances and differential states in different cases

Problem: they did not show what BH detects. That would have made for an insightful comparison

DA of microbes in infants born differently

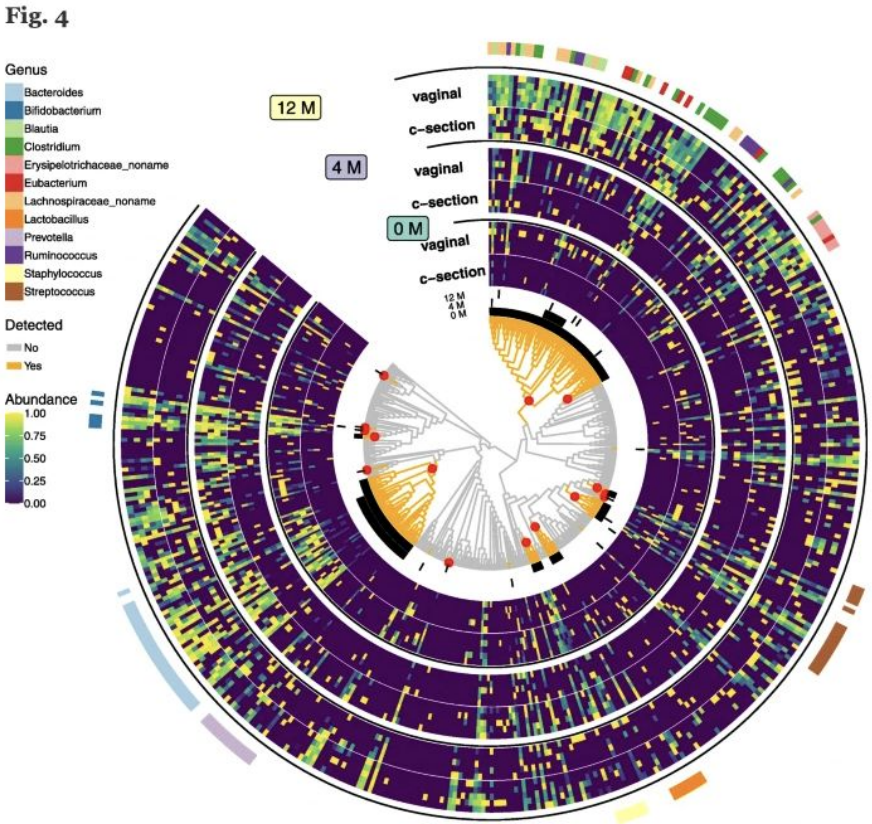
Goal: investigate whether babies born vaginally or by C-section have different microbiome compositions

Entities: gut microbiota (metaOTUs)
Data: counts metaOTUs
Tree: phylogenetic tree

DA analysis

Groups:
- Vaginal delivery
- C-section delivery

"Vaginal babies are enriched for species in genera (e.g., *Prevotella* and *Lactobacillus*) that resemble their mother's vaginal microbiota, whereas C-section newborns tend to have higher abundance of species in genera (e.g., *Staphylococcus*) that are likely to be acquired from the hospital environment or from the mother's skin."



miRNA expression analysis of cardiac pressure

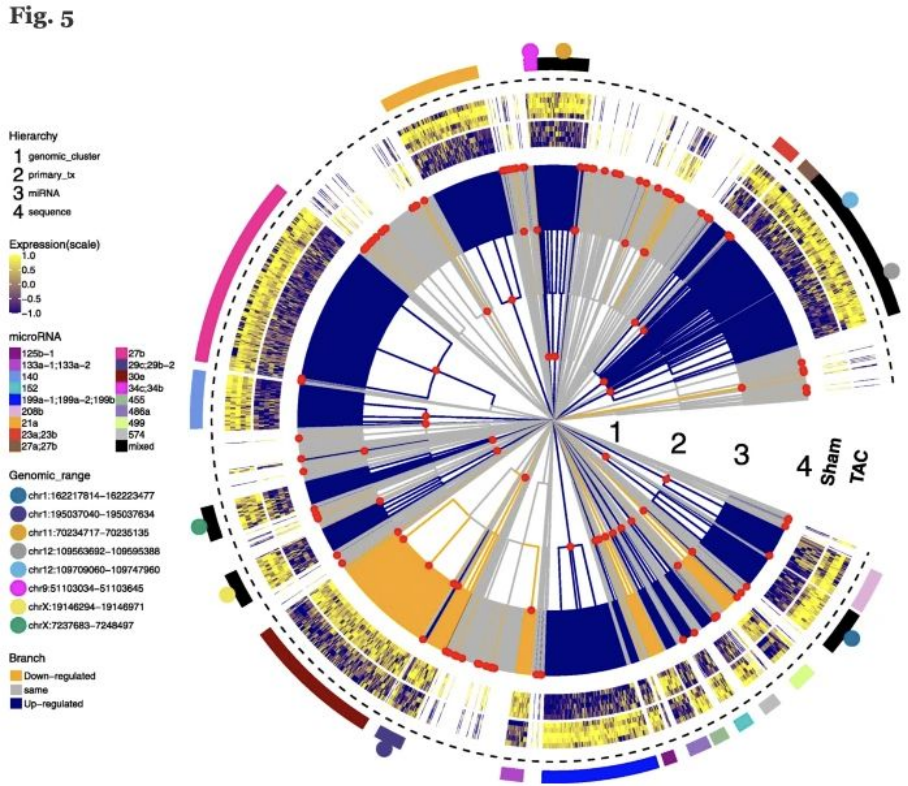
Goal: investigate whether miRNAs with the same origin are differentially co-expressed between mice receiving transaortic constriction (TAC) or mice receiving sham surgery (Sham)

Entities: miRNAs
Data: expression level
Tree: constructed based on the origin of miRNAs

DA analysis

Groups:
- Sham-treated mice
- TAC-treated mice

"While many of the identified miRNAs had previously been reported in relation to cardiovascular health and disease, our analysis highlights that most of the alterations in miRNA abundance is transcriptional, including the transcriptional co-regulation of genomic clusters containing mixed miRNA families, suggesting a common reshaping of chromatin at these regions."



DS analysis of mouse cortex scRNAseq data

Goal: understand how peripheral lipopolysaccharide (LPS) affects the mouse brain cortex

Entities: cells in mouse brain cortex
Data: expression levels of multiple genes
Tree: hierarchy of cell subpopulations

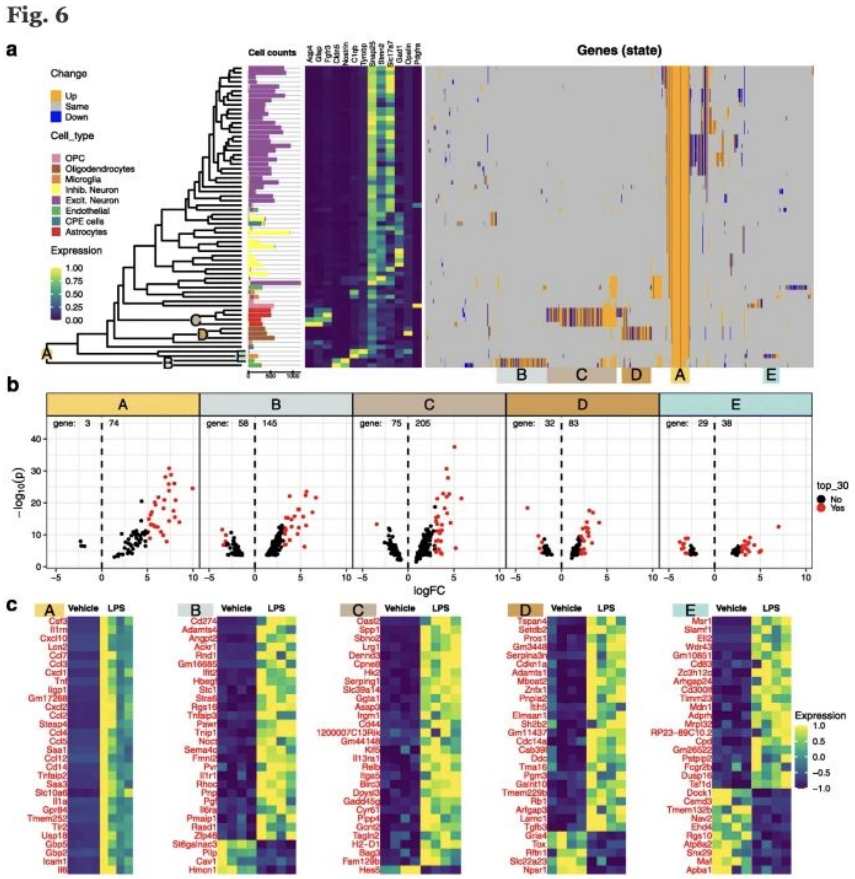
- constructed by hierarchical clustering of cells from marker genes → double-dipping of data?

DS analysis

Groups:

- Control mice
- LPS-treated mice

"Inflammatory signaling has been shown to trigger the upregulation of several cytokines in astrocytes [36], and indeed, we observe the upregulation of a number of them, including Cxcl2, Cxcl1, and Ccl5, not only in astrocytes but across all cell types (category A)."



Conclusion (according to authors)

treeclimbR is:

1. **Better** than competing methods on synthetic data
2. **Robust** against uninformative trees
3. **Particularly bad** with correlation trees (but those trees should not be used)
4. **Flexible**, as it can be used to integrate tree information to any test as long as test assigns a direction and a p-value to every node of the tree

Discussion

1. Why the need for simulating data using parametric method?
    a. Why not sample a branch from the control data and multiply all entities in that branch by a fold change?
2. No comparison against non-tree procedure BH on real datasets
3. Obscure notation (sample *k*, score *q*, score *U* ...)
4. Can the method be generalized to tests without direction?
    a. Real value (e.g. correlations)
    b. No value (e.g. t-test)