Detecting parallel evolution of bacteria in infant gut microbiomes

EVELYN BARAJAS[1], Aina Martinez Zurita[2,3], Nandita Garud[2,3]

[1] BIG Summer Program, Institute for Quantitative and Computational Biosciences, UCLA
[2] Department of Ecology and Evolutionary Biology, UCLA
[3] Department of Human Genetics, UCLA

Abstract

Infant gut microbiomes undergo significant evolutionary changes during the first year of life, due to an initial process of bacterial colonization followed by shifts in their diet. We aimed to investigate whether parallel allele frequency changes—multiple independent occurrences of the same evolutionary change across different individuals—occur in infants' gut microbiomes. We analyzed temporally sampled data from the Backhed *et al.* 2015 dataset to determine allele frequency changes at different time points in the first year of life: birth, 4 months, and 12 months. By aggregating representative non-synonymous sites on a per-gene basis and using generalized linear models, we aimed to observe clear evidence of parallelism in these allele frequency changes. Discovering parallelism would indicate that certain bacterial genes provide adaptive benefits during early gut colonization, offering insights into microbial evolution and its impact on infant gut health

Background

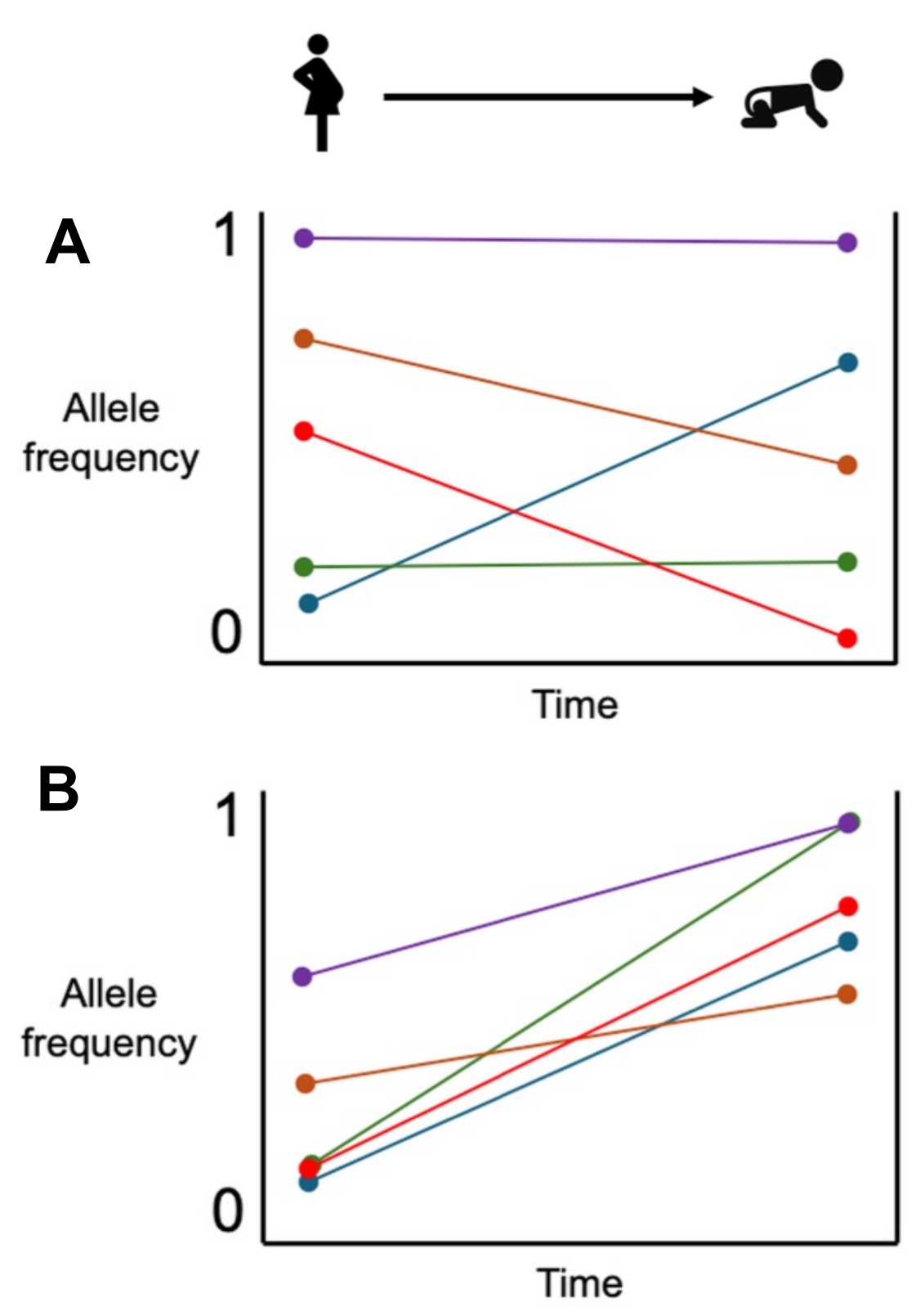

Figure 1 **A** Example of neutral state, no parallelism found. **B** Example of parallelism happening at various hosts. By Aina Martinez Zurita.

- The gut microbiome is a community of microorganisms such as bacteria obtained as early as birth.
- It plays an essential role in human health in areas such as metabolism, immunity, development, and behavior.
- Billions of mutations occur daily in the microbiome, especially during the early years of life.
- The infant gut microbiome undergoes dramatic changes during the first year, potentially due to bacterial colonization and dietary transitions.

Data

- We used Backhed *et al.* 2015 dataset, temporally sampled data of 98 mothers and their infants, to study the evolutionary changes in infant gut microbiomes and observe gene mutations happening.

Methods

Objective

Find evidence of parallelism, as it would indicate that there is positive adaptation in genes.

Initial Steps:

1. Identify the distribution of species across infants, to then identify the most prevalent species at each time point (Fig 2).
2. Identify adaptive sites that alter the amino acid of proteins and show adaptive changes.

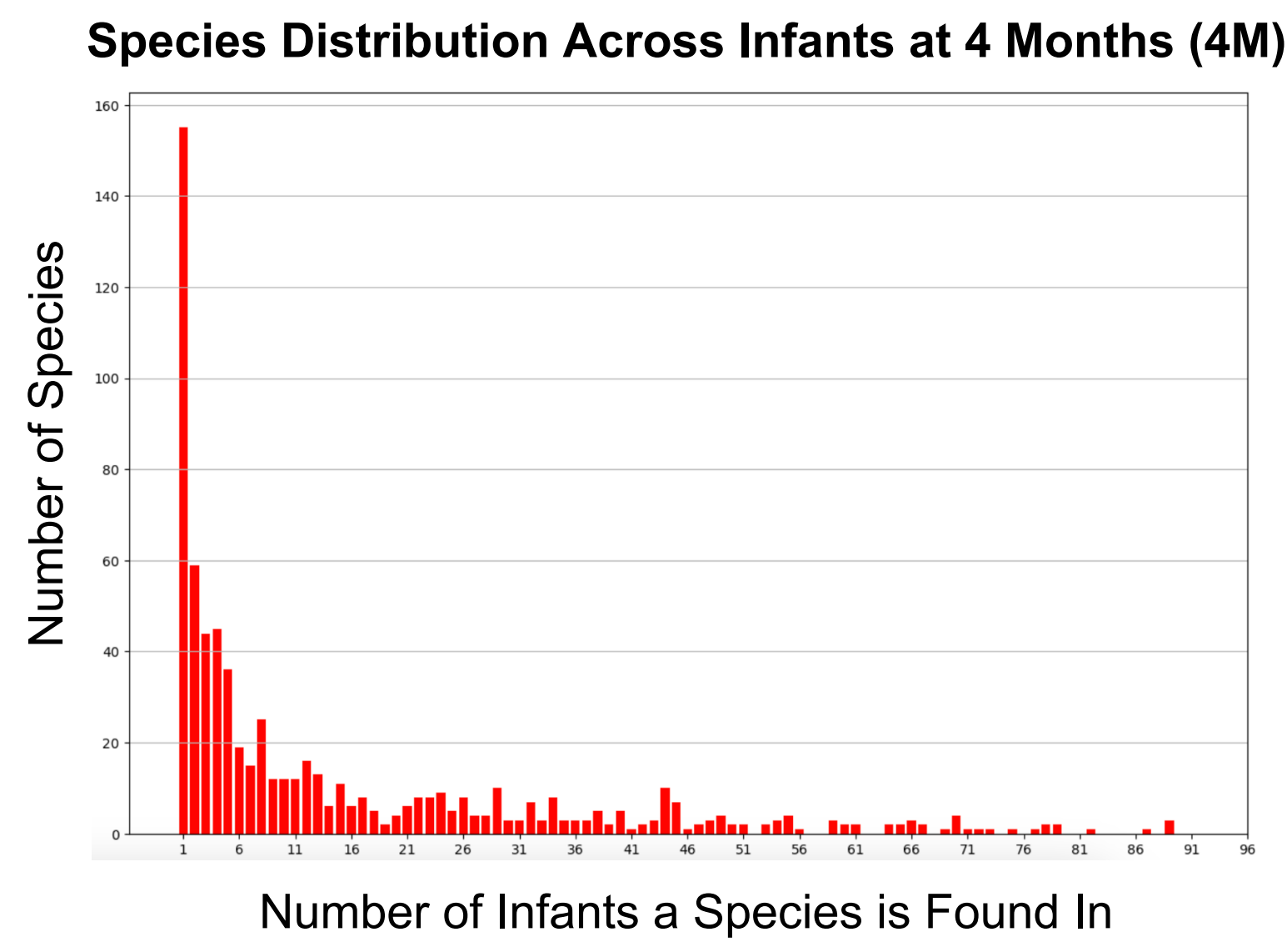

Figure 2 Species distribution across infants at 4 months.

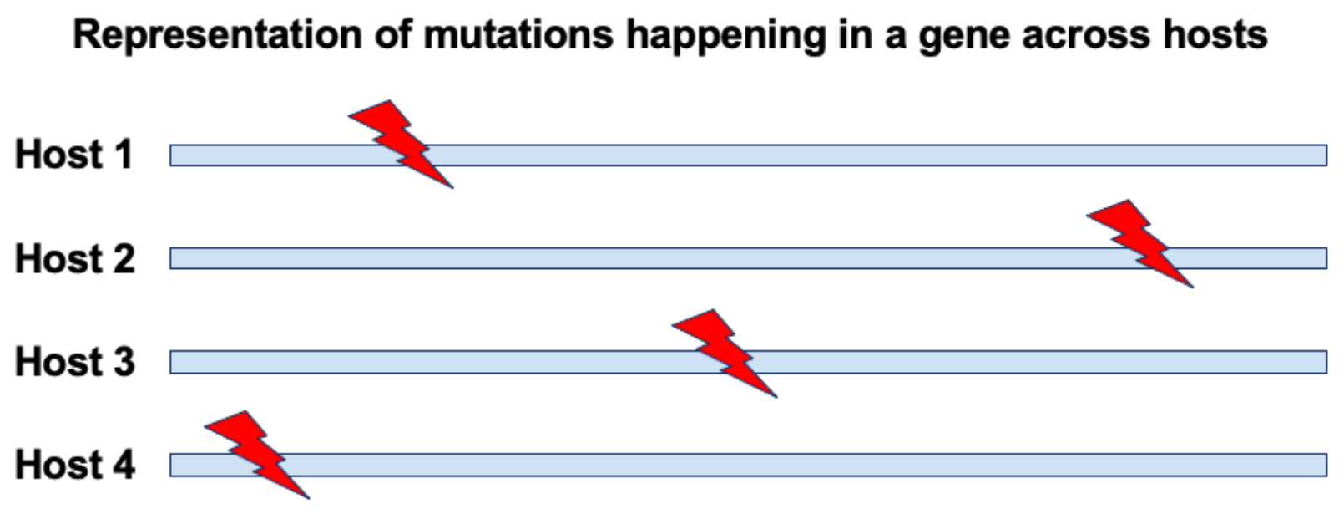
Representation of mutations happening in a gene across hosts
Host 1
Host 2
Host 3
Host 4

Figure 3 Multiple non-synonymous sites can have the same effect on the protein, regardless of their position.

Non-synonymous sites

- Non-synonymous sites (1D) have allele substitutions that change a protein's amino acid.
- Find a way to aggregate allele changes on a per-gene basis to better detect parallelism
- Great abundance of 1D sites in genes and showed the significant allele frequency changes (Fig 4).

Gene Aggregation

- Focused on Bacteroides_vulgatus_57955.
- We aggregated non-synonymous site data to examine their mutations across multiple hosts.
- Using Python, we selected and aggregated the 1D site with largest allele frequency change for each host on a per-gene basis.

GLM Analysis

- Generalized linear model was used to further examine the aggregated data.

Q-Q Plot

- Shuffled data to assess the significance of our data
- Distribution of the real data is more dispersed, with more extreme values than the shuffled data (Fig 6)

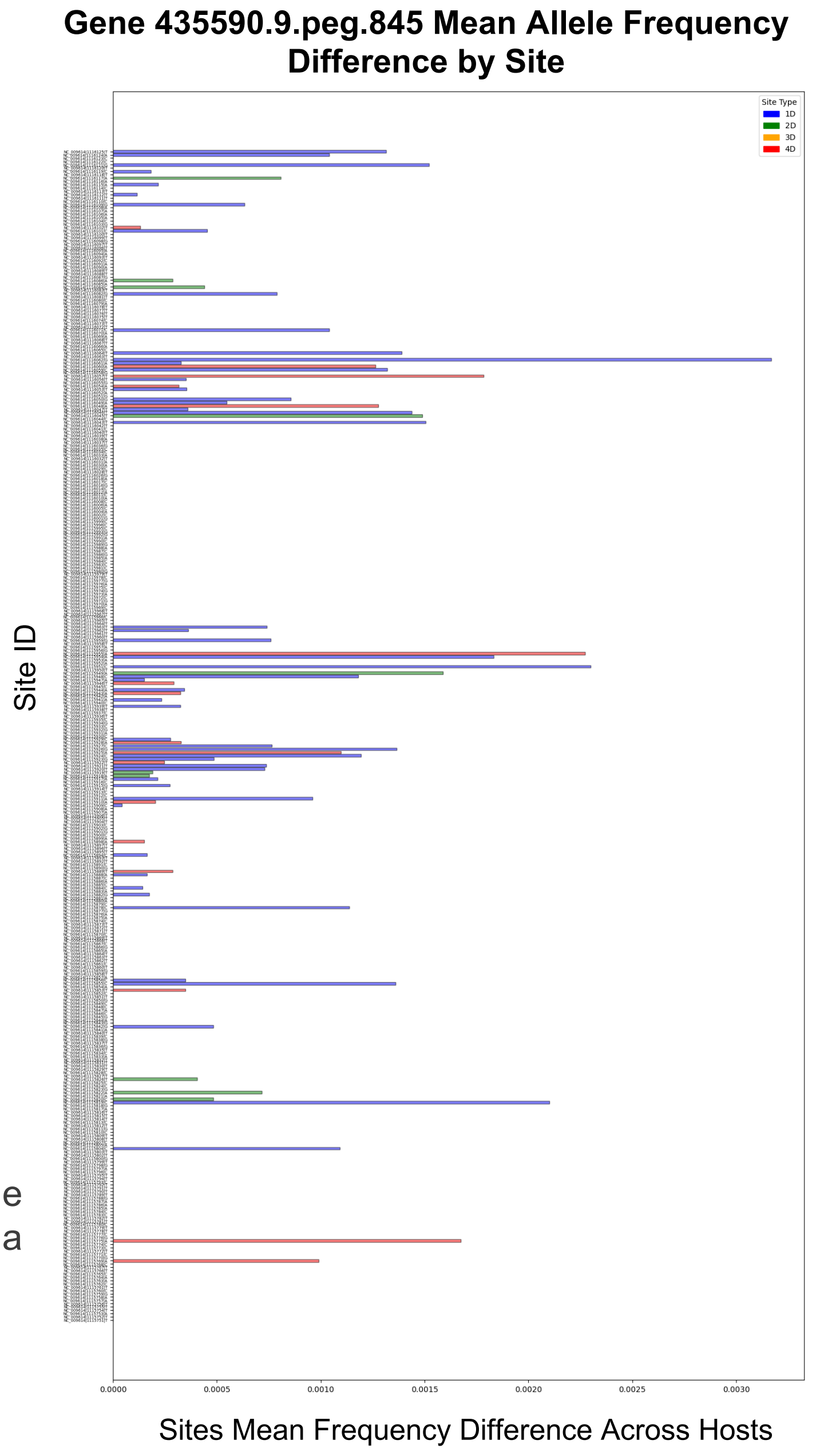

Figure 4 Gene representation for the average allele frequency change in each site that passed a read filter >50.

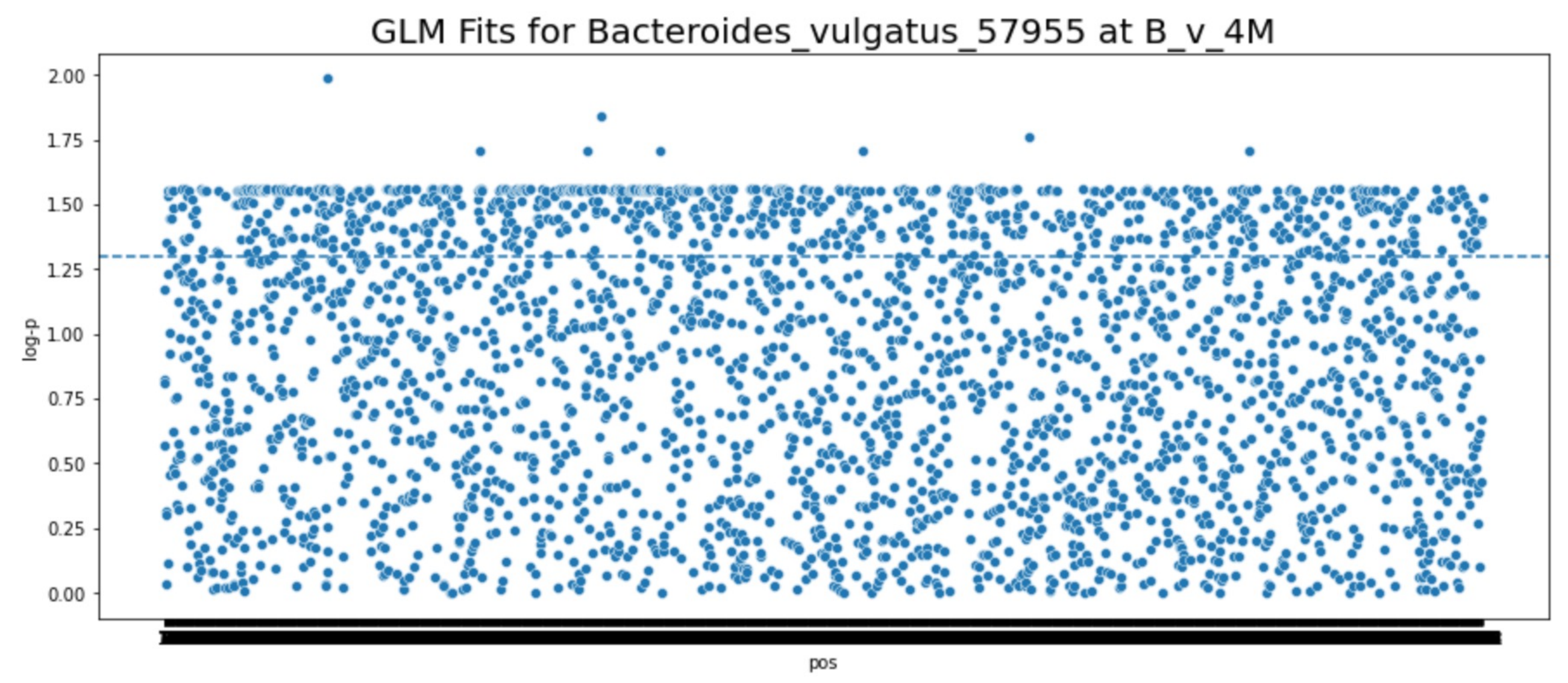

Figure 5 Identifies genes of interest based on their statistical significance in the GLM analysis. Genes with log-p values above threshold line are significant. By Aina Martinez Zurita.

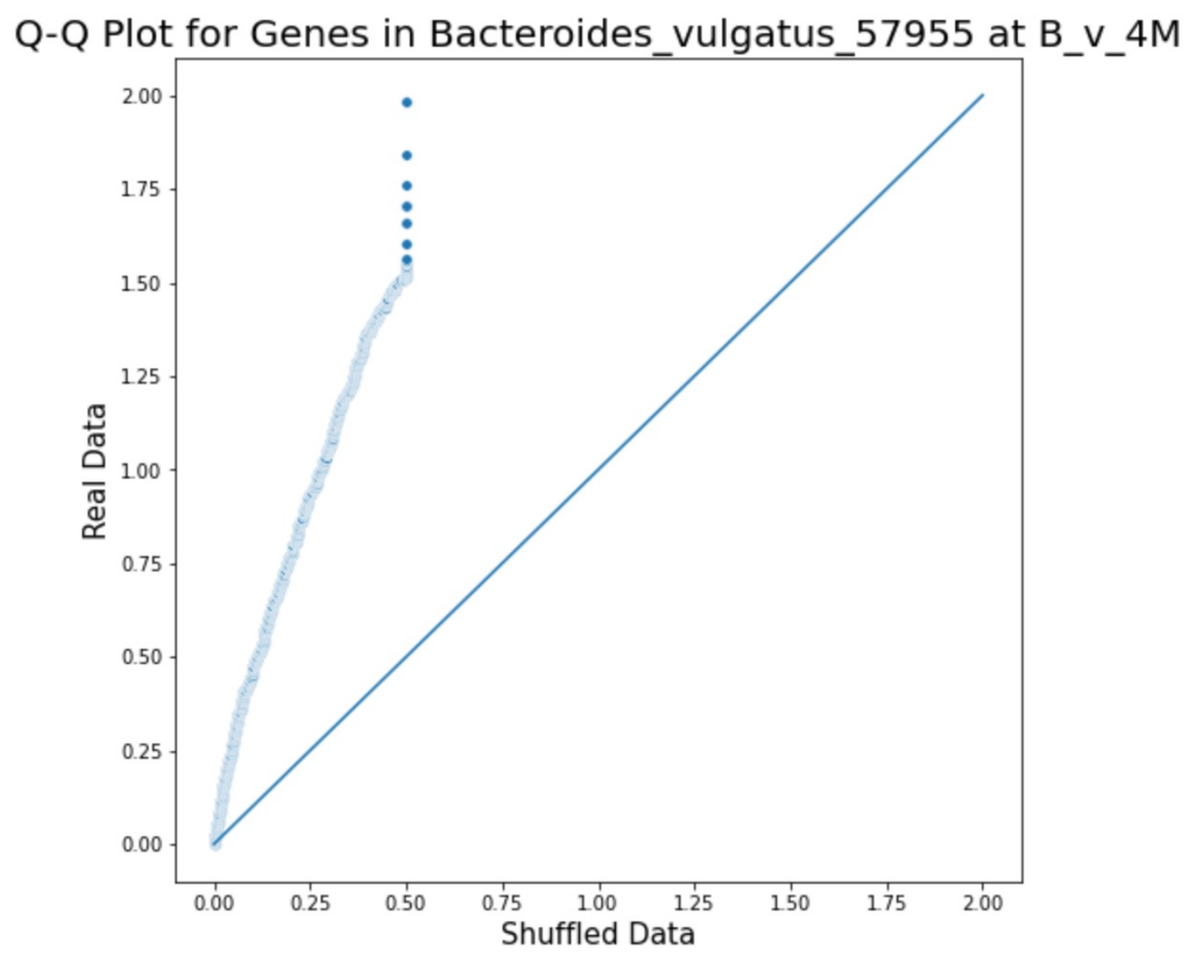

Figure 6 Distribution of real data against shuffled data for genes in Bacteroides vulgatus. By Aina Martinez Zurita

Results

Evidence of parallelism in some genes

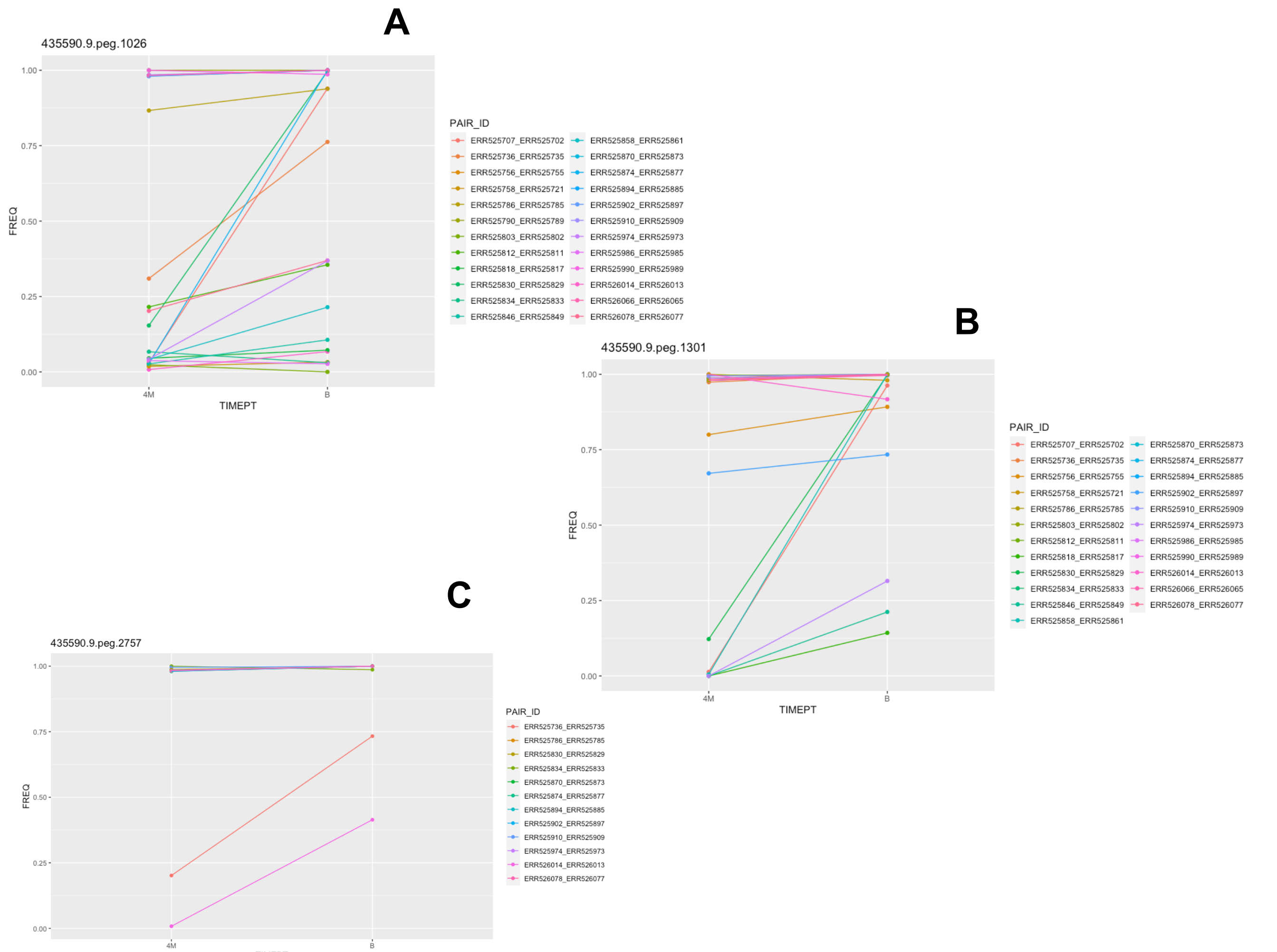

Figure 7 **A** Gene 435590.9.peg.1028 has data for 24 hosts and there are signs of parallelism. **B** Gene 435590.9.peg.1301 has data for 23 hosts. **C** Gene 435590.9.peg.2757 has data for 12 hosts. By Aina Martinez Zurita

- Some genes show putative evidence of parallelism
- Parallelism is indicated by lines that move in a similar direction and manner from between time points
- This pattern could be indicative of adaptation or other evolutionary processes acting on the allele across the different hosts.

Conclusions

- By aggregating non-synonymous sites with the largest allele frequency difference between birth and 4 months on a per-gene basis, **we found putative evidence for parallelism in some genes**.
- Some more specific validations might be required to confirm if these changes are evolutionary trends.

References

Backhed *et al.* 2015. Dynamics and stabilization of the human gut microbiome during the first year of life. *Journal of Clinical Investigation*, 17(5), 690-703.

Acknowledgments

I would like to thank my mentors Aina Martinez Zurita and Nandita Garud for their guidance, encouragement, and unwavering support throughout this project. As well as members of the Garud Lab as their kindness and expertise made this journey both enlightening and rewarding. Thank you for giving me the opportunity to work with you during the summer.