Plant Dietary Diversity and Gut Microbial Ecology in the American Gut Project: A Compositional, Constrained-Permutation, and Transportability Analysis Supporting a Redesigned Controlled-Feeding Trial

Abstract

Background. The observation from the American Gut Project (AGP) that people eating more than 30 plant types per week harbor distinct gut microbiota from those eating 10 or fewer has become an influential dietary heuristic, yet it rests on a small cross-sectional subgroup and has not been re-examined with modern compositional statistics or tested for geographic transportability. We asked whether the ">30 versus ≤ 10 plant types per week" contrast is *robustly* associated with gut microbial ecology after adjustment for antibiotics, age, body-mass index (BMI), geography and other dietary variables, and whether any association would justify funding a controlled-feeding study. **Methods.** Using the fixed 2018 AGP fecal sub-operational-taxonomic-unit (sOTU) table and mapping file (Figshare `6137192` and `6137315`), we analyzed one stool sample per participant (2676 participants; 21 640 sOTUs). We performed a missingness and selection-bias analysis, computed centered-log-ratio (CLR)/Aitchison beta-diversity, tested the exposure with partial distance-based redundancy analysis (dbRDA/PERMANOVA) using country-constrained permutations, identified a supervised compositional balance by forward selection, validated it with leave-one-country-out (LOCO) cross-validation, and derived power/sample-size requirements for a future trial. We drew no functional or metabolic inferences from 16S data. **Results.** Community-wide, the exposure explained a statistically significant but negligible fraction of Aitchison beta-diversity (Model A: pseudo-$F = 1.62$, $R^2 = 0.12\%$, constrained-permutation $p = 0.004$). A prespecified eight-genus balance separated groups strongly in-sample (adjusted coefficient $+0.450$, $t = 6.89$, $p = 8.3 \times 10^{-12}$, partial $R^2 = 3.4\%$; Cohen's $d = 0.602$) and was robust to BMI adjustment ($p = 3.3 \times 10^{-9}$). However, in LOCO validation the held-out exposure coefficient collapsed to ≈ 0 (mean -0.000, SD 0.139; 0/3 folds $p < 0.05$) versus a training mean of $+0.631$, and the balance added only -0.019 area-under-the-curve (AUC) over dietary self-report covariates. Missingness was substantial (BMI $\sim$51%) and strongly geographically structured, violating the missing-completely-at-random assumption. **Conclusions.** The endpoint is real and highly significant in-sample but does not transport across geography, indicating residual confounding rather than a portable biological effect. We therefore recommend **REDESIGN**: fund a geographically stratified, within-subject crossover feeding study powered on a conservative (transportability-adjusted) effect size—approximately 54 subjects at between-period correlation $r = 0.7$ versus 175 per group ($\sim$350 total) for a parallel design—with a single prespecified taxonomic balance endpoint.

Keywords: gut microbiome; plant diversity; compositional data analysis; Aitchison distance; PERMANOVA; transportability; American Gut Project; controlled-feeding trial; crossover design

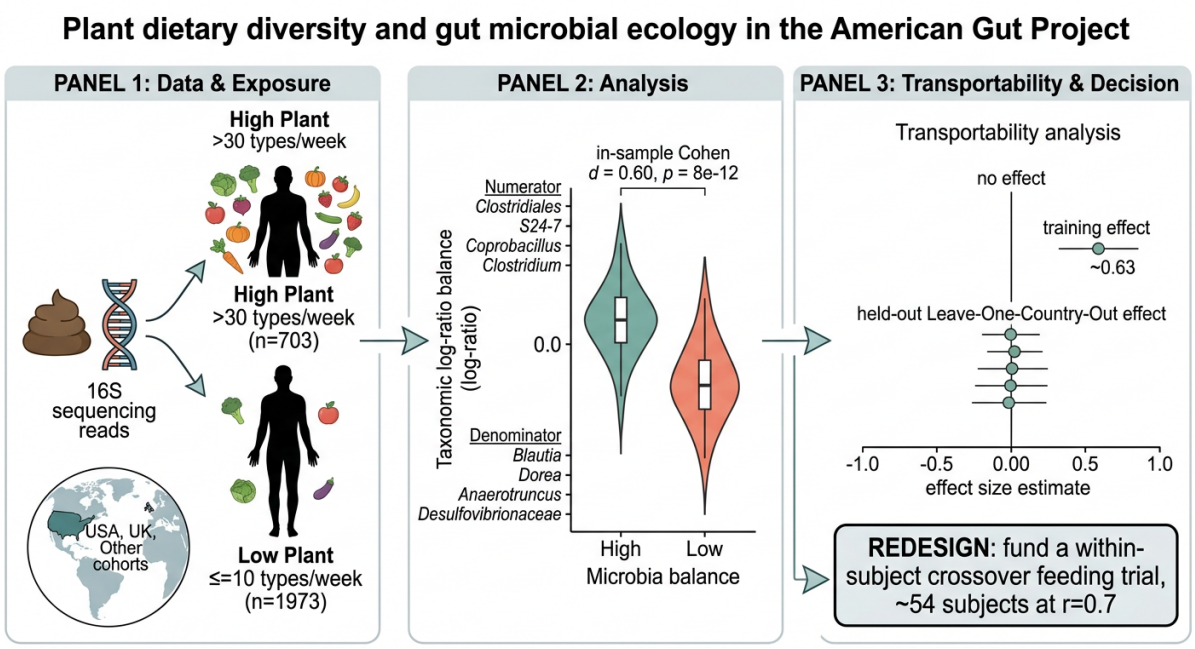


Figure 1: Graphical abstract. From the fixed American Gut Project fecal sOTU table, one stool sample per participant was classified as High Plant (>30 types/week, $n = 703$) or Low Plant (≤10 types/week, $n = 1{,}973$). A supervised taxonomic log-ratio balance separated the groups strongly in-sample (Cohen's $d = 0.60$; $p = 8 \times 10^{-12}$), but leave-one-country-out validation showed the held-out effect collapsing to ≈ 0, motivating a **REDESIGN** decision toward a within-subject crossover feeding trial (∼54 subjects at $r = 0.7$).

1 Introduction

The adult human gastrointestinal tract harbors a microbial community of a magnitude comparable to the number of human cells in the body, and this community is now recognized as a central mediator between diet and host physiology [1–3]. Large population surveys have repeatedly shown that host and environmental variables—rather than any single "healthy" configuration—explain the bulk of interindividual variation in gut community composition, with diet, medication, transit time and geography featuring prominently [4–6]. Even the notion of a shared "core" microbiome is dominated by interindividual variability, as early twin studies made clear [7]. Among dietary exposures, the quantity and, increasingly, the *diversity* of plant intake have attracted particular attention because plant cell walls supply the microbiota-accessible carbohydrates (MACs) that resident fermenters convert into short-chain fatty acids, the metabolites most strongly implicated in colonic and systemic host signalling [8–10]. Long-term dietary patterns are associated with reproducible differences in community structure [11], and controlled interventions demonstrate that switching between plant- and animal-based diets can reconfigure the community within a single day [12].

Against this backdrop, the American Gut Project (AGP)—an open, citizen-science platform that has profiled tens of thousands of self-collected stool samples by 16S ribosomal RNA (rRNA) gene amplicon sequencing—reported one of the most widely propagated dietary heuristics in the field: participants who reported consuming more than 30 different plant types per week displayed greater microbial α-diversity and a distinct community structure relative to those consuming 10 or fewer, apparently independent of whether they identified as omnivore or vegan [13]. This "30 plants

per week" rule of thumb has since been adopted in the popular and clinical nutrition literature as a practical target for a diverse, fiber-rich diet [14, 15]. Yet the original comparison rested on a small subgroup (on the order of tens of participants per arm) nested within a much larger, self-selected cohort, and it has not been re-examined with the statistical machinery that the microbiome field now regards as mandatory.

Two methodological developments make such a re-examination timely. First, sequence-count data are *compositional*: they carry only relative information constrained to a simplex, so that analyses performed on raw proportions or rarefied counts can generate spurious associations and are, in a formal sense, inadmissible [16–18]. Log-ratio transformations—the centered log-ratio (CLR) underlying the Aitchison distance, isometric log-ratio coordinates, and supervised "balances" between groups of taxa—respect this geometry and yield interpretable, scale-invariant contrasts [19–22]. Second, the field has become acutely aware that microbiome associations frequently fail to generalize. Host variables can confound disease signals [23]; technical and batch variation from collection, extraction and sequencing can rival biological signal [24, 25]; and geography exerts so strong an influence that models trained in one region often perform poorly in another [26–28]. High within-cohort performance that collapses under external or cross-cohort validation is a recurring and cautionary pattern, and reporting standards such as STORMS now explicitly ask authors to document confounding and validation strategy [29, 30].

These considerations reframe the practical question facing a funder. It is not merely whether plant diversity is *associated* with gut microbial ecology in the AGP—an observational, self-reported exposure embedded in a geographically structured, heavily missing dataset—but whether any association is robust and *transportable* enough to justify the expense of a controlled-feeding study, and if so, how that study should be designed. In this work we address that question directly. Using the fixed 2018 AGP fecal sOTU table and mapping file, we (i) characterize missingness and selection bias in the exposure-defining cohort; (ii) quantify the community-wide association with compositional (Aitchison) beta-diversity using constrained-permutation partial dbRDA/PERMANOVA that holds geography fixed; (iii) identify a supervised, prespecifiable compositional balance that best separates high- from low-plant consumers after covariate adjustment; (iv) subject that balance to leave-one-country-out (LOCO) validation to test transportability; and (v) translate the resulting effect sizes into formal power and sample-size requirements for a future trial. We deliver adjusted effects, balance and transportability diagnostics, and an explicit *fund/redesign/do-not-fund* decision together with a single prespecified microbial endpoint. Throughout, we honor a strict inferential constraint: because 16S amplicon data measure taxonomy rather than gene content or activity, we make *no* functional or metabolic-pathway claims and confine every endpoint to compositional, taxonomic terms [31, 32].

2 Methods

2.1 Data acquisition and provenance

Two fixed public artifacts of the American Gut Project were obtained programmatically through the Figshare REST API: the fecal sOTU count table (article `6137192`; `deblur_125nt_no_blooms.biom`, 25 MB), generated by the Deblur denoising workflow on 125-nucleotide 16S reads with bloom sequences removed [33]—a sub-OTU resolution comparable to the amplicon-sequence-variant approach of DADA2 [34] and consistent with the reproducible QIIME 2 ecosystem [35]—and the full AGP mapping file (article `6137315`; `corrected-t2.tsv`, 109 MB) [13]. Downloads were checksum-verified against the API-reported MD5 values. Taxonomy followed the Greengenes reference used by AGP

[36]. All processing used Python 3.12 with a fixed random seed (42); the complete pipeline is scripted and versioned.

2.2 Cohort definition and exposure

The mapping file (17 854 records) was restricted to samples with sequencing data present in the BIOM table (9511 fecal samples). Every fecal sample carried a valid and unique host-subject identifier, so the table already represented one fecal sample per participant; a deduplication step that would otherwise retain the most metadata-complete sample per participant was confirmed to be a no-op. The exposure was defined from the `types_of_plants` questionnaire item: **High Plant** (>30 types/week) comprised the "More than 30" response ($n = 703$); **Low Plant** ($\leq$10 types/week) pooled "6 to 10" and "Less than 5" ($n = 1973$). Intermediate consumers ("11 to 20" and "21 to 30"; $n = 3368$) and non-responders ($n = 3467$) were excluded to sharpen the contrast. Aligning the sOTU table to the cohort and dropping all-zero features yielded a final analytic matrix of 21 640 sOTUs by 2676 participants (Figure 2).

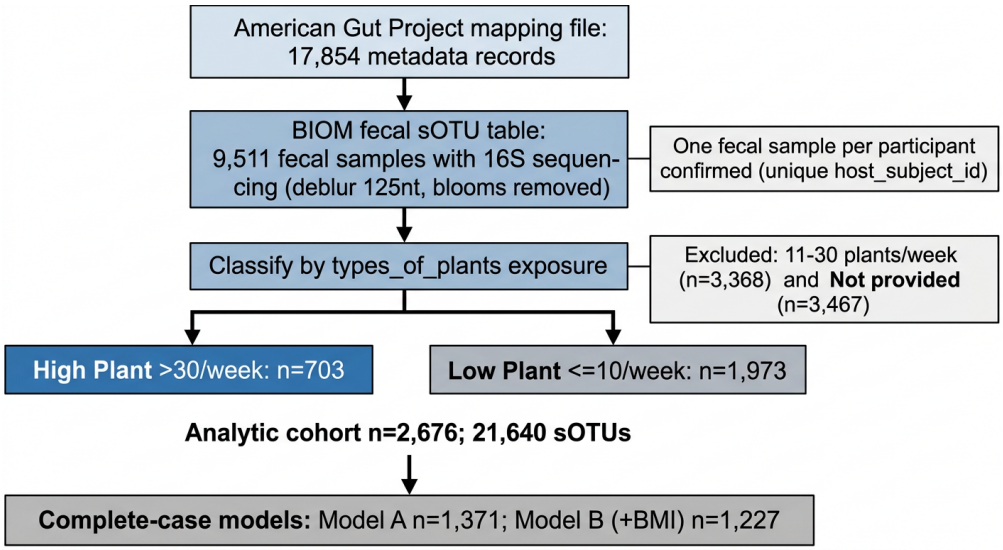


Figure 2: Participant and sample selection. Flow from the AGP mapping file and fecal sOTU BIOM table to the analytic cohort and the complete-case regression models. One fecal sample per participant was confirmed by unique host-subject identifier; intermediate (11–30 plants/week) and non-responding participants were excluded to define the High- versus Low-Plant contrast.

Covariates and a documented data limitation

Adjustment covariates comprised age, sex, BMI (`bmi_corrected`), country/geography, antibiotic history, and three additional dietary-frequency items (alcohol, meat/eggs, and vegetable frequency). Geography was pooled to three levels (USA, United Kingdom, Other) to stabilize estimation given the strongly USA-weighted cohort. We record one unavoidable limitation: the AGP Figshare mapping file contains *no* sequencing-plate or run identifier column, so sequencing plate—nominally a target adjustment variable and a canonical source of batch effect [24, 25]—is 100% missing and could neither be summarized nor modelled. Geography, which is highly collinear with AGP sample-processing logistics, serves as the principal available proxy for batch and population structure.

2.3 Missingness and selection-bias analysis

A panel of missingness tokens (e.g., "not provided", "unknown", "not collected") was harmonized to a single missing state. We tabulated per-covariate missingness overall and by exposure group, and compared complete cases (all key covariates present) with incomplete cases on age (Mann–Whitney U and Welch t), sex and pooled country (χ^2), and exposure group (χ^2). A logistic regression modelled the probability of being a complete case as a function of exposure, standardized age, sex and pooled country, and the missing-completely-at-random (MCAR) assumption was assessed from the univariate comparisons.

2.4 Compositional transformation and Aitchison beta-diversity

Counts were CLR-transformed (natural logarithm, pseudocount 1), so that the Euclidean distance between CLR vectors is the Aitchison distance [16, 17]. Principal-component analysis of the CLR matrix (Aitchison PCA) provided an unsupervised ordination with 95% confidence ellipses by exposure group; unlike phylogenetic distances such as UniFrac [37], the Aitchison distance is a purely compositional metric that requires no reference tree.

2.5 Constrained-permutation dbRDA / PERMANOVA

The exposure was tested against Aitchison beta-diversity using partial distance-based redundancy analysis with a pseudo-F statistic, equivalent in this metric to PERMANOVA [38, 39]. Covariate effects were removed by Frisch–Waugh–Lovell residualization (conditioning matrix), and the partial, adjusted R^2 was computed in the `vegan` style as $R^2_{\text{adj}}(\text{full}) - R^2_{\text{adj}}(\text{conditioned})$. Crucially, statistical significance was assessed by permuting the exposure labels *within* pooled-country strata (999 permutations), so that the null distribution preserves geographic structure and the test cannot be driven by between-country differences [40]. Two nested models were fitted: Model A (age, sex, country, antibiotics, alcohol, meat/eggs and vegetable frequency; complete-case $n = 1371$) and Model B (Model A + BMI; $n = 1227$).

2.6 Supervised compositional balance selection

To move from a negligible community-wide signal to a prespecifiable, targeted endpoint, sOTUs were aggregated to genus-level Greengenes lineages, filtered to genera present in $\geq$10% of samples, and reduced to the top 100 by prevalence (86 genera considered after filtering). A "balance" was defined as

$$\mathrm{B} = \frac{1}{|N|}\sum_{i\in N} \ln x_i \;-\; \frac{1}{|D|}\sum_{j\in D} \ln x_j, \tag{1}$$

the difference in mean log-abundance between a numerator taxon set N and a denominator set D (pseudocount 0.5), following the selbal balance framework [20, 21]. Forward selection (maximum four taxa per side) added the taxon that most increased the absolute t-statistic of the exposure term in a Model-A–adjusted ordinary-least-squares regression of B, with a minimum relative-t gain of 0.02 as the stopping rule. Robustness was assessed by (i) refitting with BMI (Model B), (ii) 5-fold stratified cross-validation in which the entire selection was re-run within each training fold, and (iii) tabulating taxon selection frequency across folds. The in-sample standardized effect size (Cohen's d) was computed from the group means and pooled standard deviation of B.

2.7 Leave-one-country-out (LOCO) transportability validation

Transportability was tested by holding out each pooled-country cohort (USA, United Kingdom, Other) in turn. Within the training countries only, the top-genus feature space was fixed, the balance was *re-selected* de novo, and a Model-A regression (with country omitted, since a single country was held out) was fitted; the resulting balance and coefficients were then applied to the held-out country. We report, per fold, the training and held-out exposure coefficients with 95% confidence intervals, the out-of-sample area under the receiver-operating-characteristic curve (AUC) of a logistic classifier predicting High-Plant status, and the Jaccard overlap between the fold-selected and globally selected taxa. To attribute discriminative signal, we compared three classifiers: the full model (balance + all covariates), a covariates-only model (no balance), and a balance-plus-demographics model (dietary self-report removed). The marginal contribution of the balance was the AUC difference between the full and covariates-only models.

2.8 Power and sample-size analysis

Sample sizes for a future controlled-feeding study were computed with two-sided $\alpha = 0.05$ at 80% and 90% power using `statsmodels` power solvers [41, 42]. For a parallel-group (independent two-sample t-test) design we used both the in-sample $d = 0.602$ and a deliberately conservative, transportability-adjusted $d = 0.301$ (a 0.5× shrinkage reflecting the LOCO collapse). For a within-subject crossover (paired t-test) design we converted the between-subject d to a paired effect size $d_z = d/\sqrt{2(1-r)}$ for between-period correlations $r = 0.5$ and $r = 0.7$. The prespecified primary endpoint for the trial was fixed to the exact global balance identified here.

2.9 Software, reproducibility and inference constraint

Analyses used Python 3.12 with `numpy`, `pandas`, `scikit-learn` [43], `statsmodels` [41], `biom-format` and `h5py`; all steps used random seed 42 and are provided as versioned scripts with JSON result logs. Consistent with the study's core constraint, all endpoints are purely taxonomic/compositional; no functional, enzymatic or metabolic-pathway activity was inferred from 16S data, and predictive functional-profiling tools were deliberately *not* applied [31, 32].

3 Results

3.1 Cohort characteristics

The analytic cohort comprised 2676 participants—703 High Plant and 1973 Low Plant—profiled at 21 640 sOTUs (Table 1). The cohort was middle-aged (overall mean age 45.1 years), predominantly from the USA (2023/2676, 75.6%) with the United Kingdom the next largest stratum (454, 17.0%), and the majority reported no antibiotic use in the prior year. High-Plant participants had a modestly lower mean BMI (23.4 versus 24.4 $\mathrm{kg/m^2}$) and a higher proportion of female participants, foreshadowing the covariate imbalance that motivates adjustment.

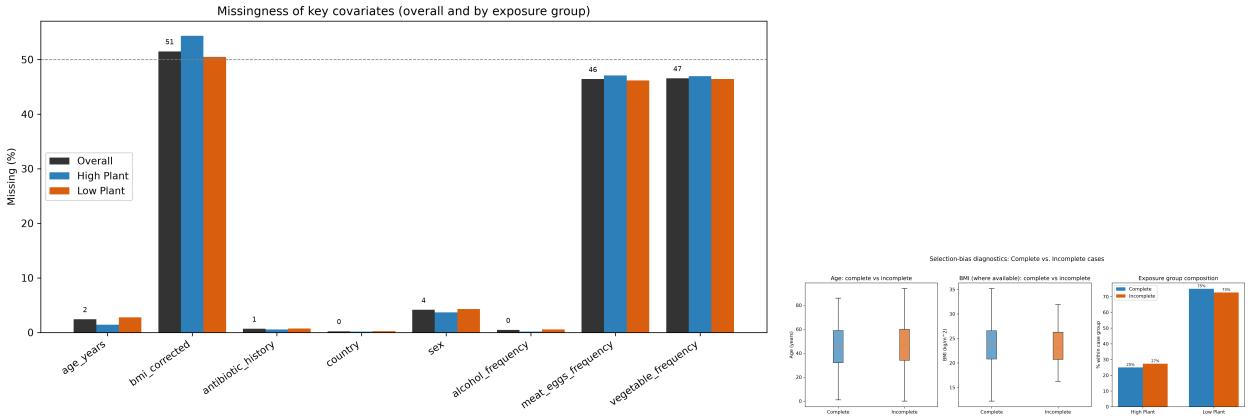
Table 1: Analytic cohort characteristics by plant-consumption exposure. Values are means (SD) for continuous variables or counts (%) for categorical variables; BMI and dietary-frequency items are substantially missing (see Section 3.2).

Characteristic	High Plant (>30/wk)	Low Plant (≤10/wk)	Overall
Participants, n	703	1,973	2,676
Age, years, mean (SD)	46.4 (15.6)	44.6 (18.6)	45.1 (17.8)
BMI, kg/m^2, mean (SD)	23.4 (4.5)	24.4 (5.4)	24.2 (5.2)
BMI missing, %	54.3	50.5	51.5
Female, n	444	886	1,330
USA, n (%)	511 (72.7)	1,512 (76.6)	2,023 (75.6)
United Kingdom, n (%)	150 (21.3)	304 (15.4)	454 (17.0)
No antibiotics in past year, n	510	1,234	1,744

3.2 Missingness is substantial and geographically structured

Covariate missingness was highly heterogeneous (Figure 3a). Age (2.4% missing), antibiotic history (0.7%), country (0.2%) and alcohol frequency (0.4%) were nearly complete, but BMI was missing for 51.5% of participants and the meat/eggs and vegetable-frequency items were each missing for ∼46%. Because the adjustment set includes these diet-frequency items, the realized complete-case sample for Model A was 1371 and, after adding BMI, 1227 for Model B—less than half of the nominal cohort (45.9% complete).

Complete and incomplete cases did not differ by exposure group (χ^2 $p = 0.19$) or sex ($p = 0.85$), but they differed by age (Mann–Whitney $p = 0.038$) and, dramatically, by geography (χ^2 $p = 2.2 \times 10^{-115}$): United Kingdom participants were far more likely to be complete cases than USA participants. A logistic model of completeness confirmed this structure, with markedly elevated odds of completeness for United Kingdom (odds ratio [OR] 26.4, 95% CI 17.9–39.0) and Other (OR 3.8, 95% CI 2.8–5.3) participants relative to the USA, a small age effect (OR 0.90 per SD), and modestly *lower* completeness among High-Plant participants (OR 0.70, 95% CI 0.57–0.86). The univariate differences on age and country reject the MCAR assumption; missingness is therefore related to observed characteristics (missing-at-random at best), and a naive complete-case analysis carries a quantifiable selection-bias risk. Reassuringly, complete and incomplete cases overlapped closely on the *distributions* of age and (where available) BMI, and the exposure-group composition was nearly identical across the two (Figure 3b), indicating that the dominant missingness driver is geographic rather than exposure-linked.



(a) **(b)**

Figure 3: Missingness and selection-bias diagnostics. (a) Percentage missing for each key covariate, overall and by exposure group; BMI (∼51%) and the meat/eggs and vegetable-frequency items (∼46%) dominate, whereas age, antibiotics, country and alcohol are near-complete. (b) Complete versus incomplete cases overlap closely on age and BMI distributions and have near-identical exposure-group composition (25–27% High Plant), so the strong missingness structure is driven chiefly by geography rather than by exposure.

3.3 The community-wide association is significant but negligibly small

In the unsupervised Aitchison PCA, the first two principal components captured only 8.0% and 3.3% of total CLR variance, and the 95% confidence ellipses of the two exposure groups were almost completely superimposed (Figure 4)—the expected picture for a subtle dietary contrast embedded in a highly individualized community. Formal testing by country-constrained partial dbRDA nonetheless detected a signal. In Model A ($n = 1371$), the exposure explained $R^2 = 0.116\%$ of Aitchison beta-diversity (adjusted $R^2 = 0.045\%$) with pseudo-$F = 1.62$ and a within-country permutation $p = 0.004$. Adding BMI (Model B, $n = 1227$) left the picture essentially unchanged: $R^2 = 0.111\%$, pseudo-$F = 1.40$, $p = 0.013$. The association therefore survives adjustment and geographic permutation, but its *magnitude* at the whole-community level is on the order of one part in a thousand—far too small to serve as a trial endpoint. This dissociation between statistical significance and effect size directly motivates a targeted, prespecified balance rather than an omnibus community test.

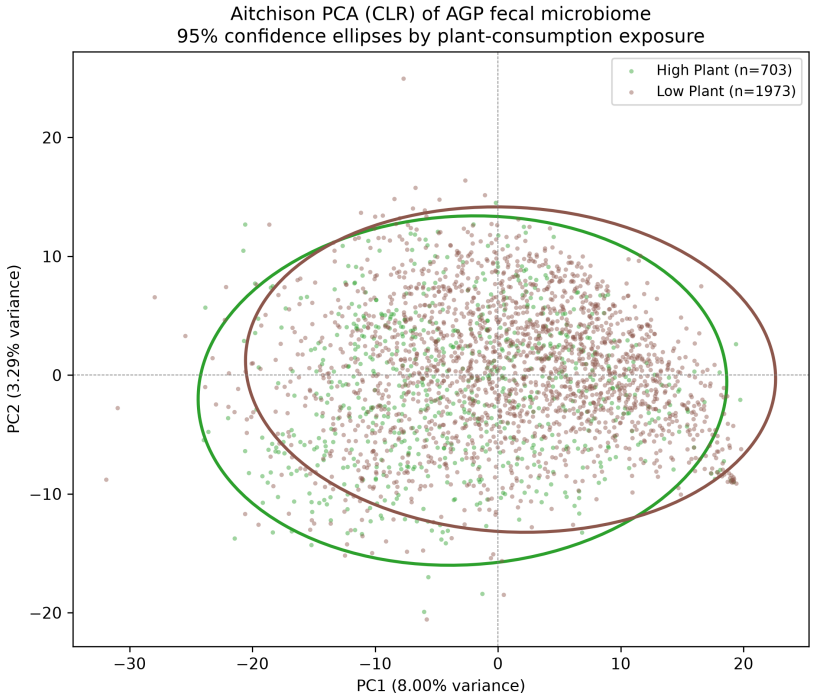


Figure 4: Aitchison (CLR) principal-component ordination of the AGP fecal microbiome with 95% confidence ellipses by exposure group. PC1 and PC2 capture 8.0% and 3.3% of total compositional variance; the near-complete overlap of the High- and Low-Plant ellipses illustrates the very small community-wide effect size, consistent with the $R^2 \approx 0.1\%$ from constrained dbRDA.

3.4 A supervised compositional balance separates the groups strongly in-sample

Forward selection identified an eight-genus balance whose numerator comprised *o__Clostridiales*, *f__S24-7*, *g__Coprobacillus* and *g__Clostridium*, and whose denominator comprised *g__Blautia*, *g__Dorea*, *g__Anaerotruncus* and *f__Desulfovibrionaceae* (positive balance → higher numerator abundance). The balance was substantially higher in High-Plant participants (mean −0.155, SD 0.955) than in Low-Plant participants (mean −0.716, SD 0.922; Figure 5), a raw difference of +0.560 that corresponds to a standardized effect of Cohen's $d = 0.602$. After Model-A adjustment, the exposure coefficient was +0.450 (95% CI 0.322–0.578; $t = 6.89$; $p = 8.3 \times 10^{-12}$), with a partial R^2 of 3.4%—roughly a thirty-fold larger share of variance than the whole-community test, precisely because the balance concentrates the signal into an interpretable log-ratio. The effect was robust to BMI: in Model B the coefficient was +0.416 (95% CI 0.279–0.553; $p = 3.3 \times 10^{-9}$; partial $R^2 = 2.9\%$), and the fully adjusted coefficient plot (Figure 6) shows the plant exposure as by far the largest and most precisely estimated term, dwarfing age, BMI, sex, antibiotic and other dietary-frequency coefficients.

Five-fold cross-validation gave an out-of-sample R^2 of 0.104 ± 0.033, and the leading numerator taxa recurred across folds—*o__Clostridiales*, *f__S24-7* and *g__Clostridium* were each selected in all five folds and *g__Anaerotruncus* in four—supporting the internal stability of the endpoint. On the strength of these in-sample diagnostics alone, the balance is a defensible, reproducible candidate primary endpoint.

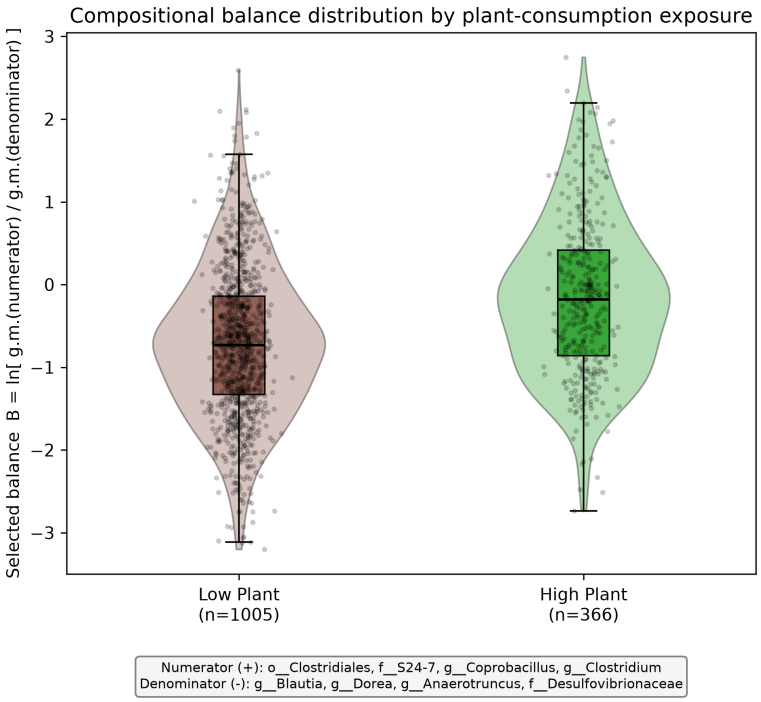


Figure 5: Distribution of the selected compositional balance by exposure group. The eight-genus log-ratio balance (numerator: *o___Clostridiales*, *f___S24-7*, *g___Coprobacillus*, *g___Clostridium*; denominator: *g___Blautia*, *g___Dorea*, *g___Anaerotruncus*, *f___Desulfovibrionaceae*) is shifted upward in High-Plant participants (mean -0.155 vs -0.716; Cohen's $d = 0.602$).

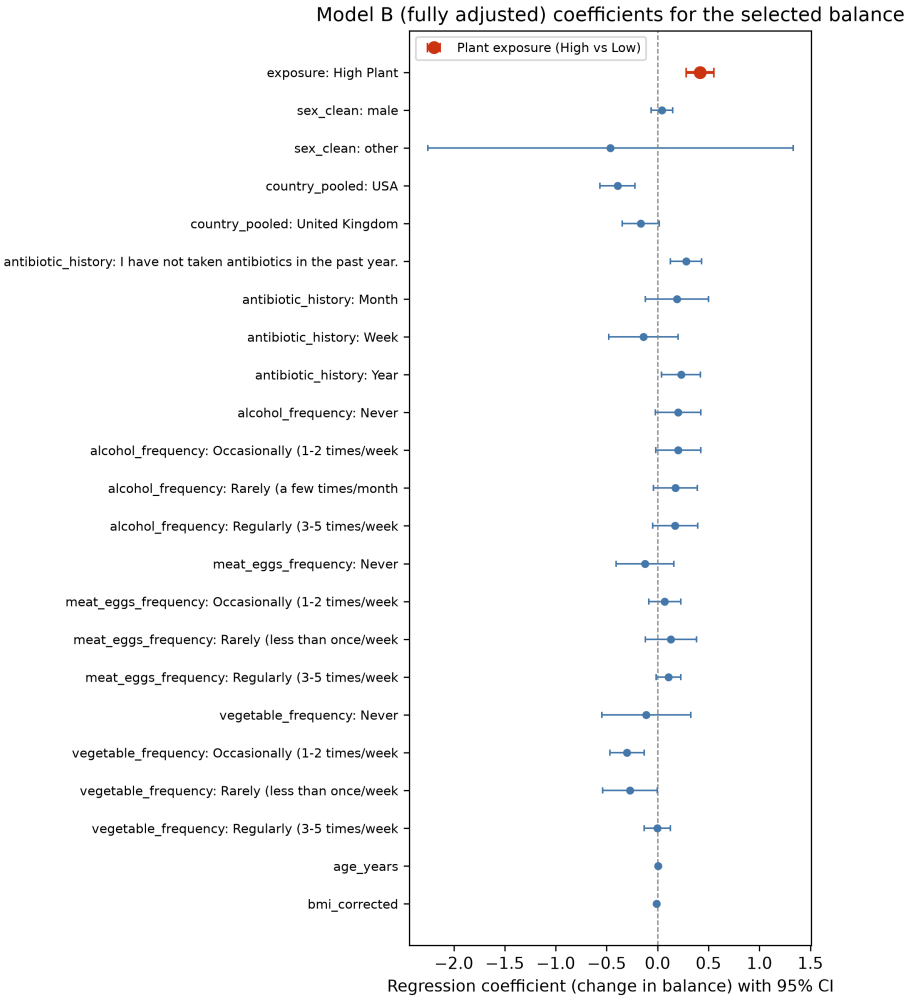


Figure 6: Fully adjusted (Model B) regression coefficients for the selected balance. The High- versus Low-Plant exposure (top, red) is the largest and most precisely estimated predictor of the balance, remaining strongly positive (+0.416; $p = 3.3 \times 10^{-9}$) after adjustment for sex, geography, antibiotic history, alcohol, meat/eggs and vegetable frequency, age and BMI.

3.5 The balance does not transport across geography

The decisive test was transportability. When each pooled-country cohort was held out and the balance was re-selected from the remaining countries, the training exposure coefficients were consistently large and significant (mean +0.631, SD 0.101; range +0.517 to +0.705; all $p < 10^{-8}$), yet the held-out coefficients collapsed toward zero (mean −0.000, SD 0.139; Figure 7). Holding out the USA gave a held-out coefficient of +0.139 (95% CI −0.072 to 0.350; $p = 0.20$); holding out the United Kingdom gave −0.000 (95% CI −0.292 to 0.292; $p = 0.999$); and holding out Other gave −0.140 (95% CI −0.593 to 0.313; $p = 0.54$). In no fold (0/3) did the held-out effect reach $p < 0.05$, and in only one fold was it even directionally consistent with the global positive effect. The mean Jaccard overlap between fold-selected and globally selected taxa was only 0.35, indicating that the *identity* of the discriminating taxa also shifted between geographic cohorts.

Out-of-sample classification told a complementary story (Figure 8). The full classifier reached a healthy mean AUC of 0.797 (range 0.788–0.814), which at first glance looks like successful

transportability. However, an attribution analysis showed this performance was carried almost entirely by dietary self-report covariates that proxy the exposure: a covariates-only classifier (no balance) already reached AUC 0.817, so adding the balance changed AUC by only −0.019 on average, whereas a balance-plus-demographics classifier with the dietary proxies removed fell to AUC 0.632. In other words, the microbial balance contributed essentially no independent out-of-sample discrimination beyond what a subject's self-reported diet already provides. Taken together—near-zero held-out regression coefficients, no marginal AUC, and unstable taxon selection—the specific linear balance is cohort- and geography-specific and does *not* transport as a fixed endpoint.

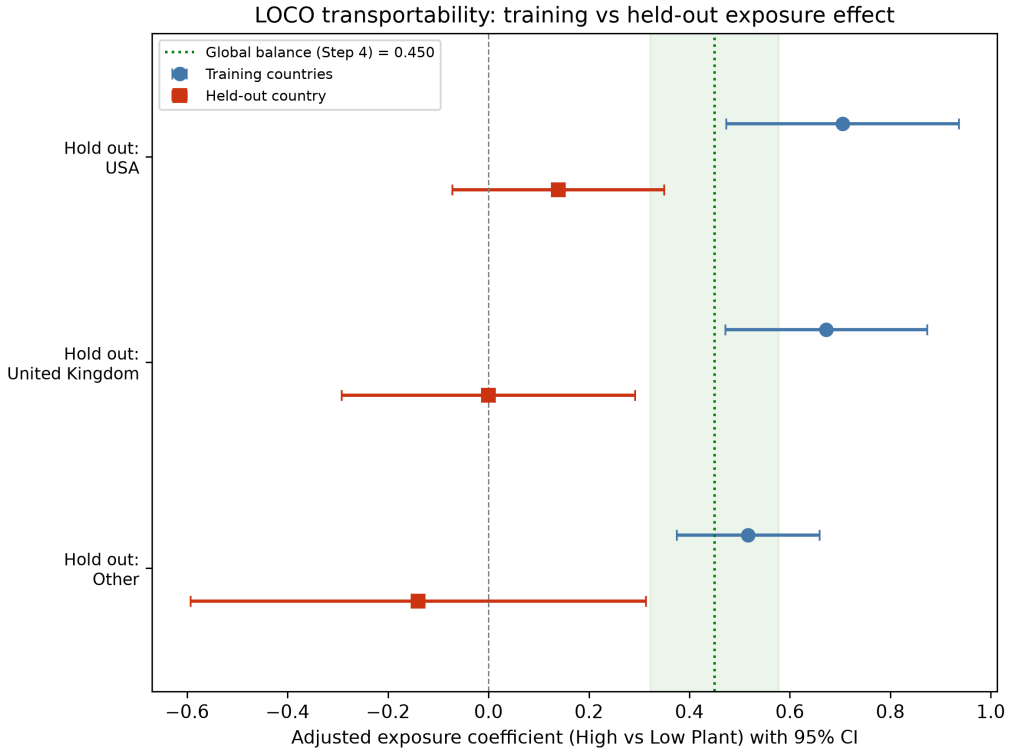


Figure 7: Leave-one-country-out transportability of the plant-exposure effect. Training-country coefficients (blue) are consistently large and cluster around the global Step-4 estimate (dotted line, +0.450; shaded 95% CI), but the corresponding held-out coefficients (red) collapse to ≈ 0 with wide intervals crossing the null in every fold, demonstrating failure to transport across the USA, United Kingdom and Other cohorts.

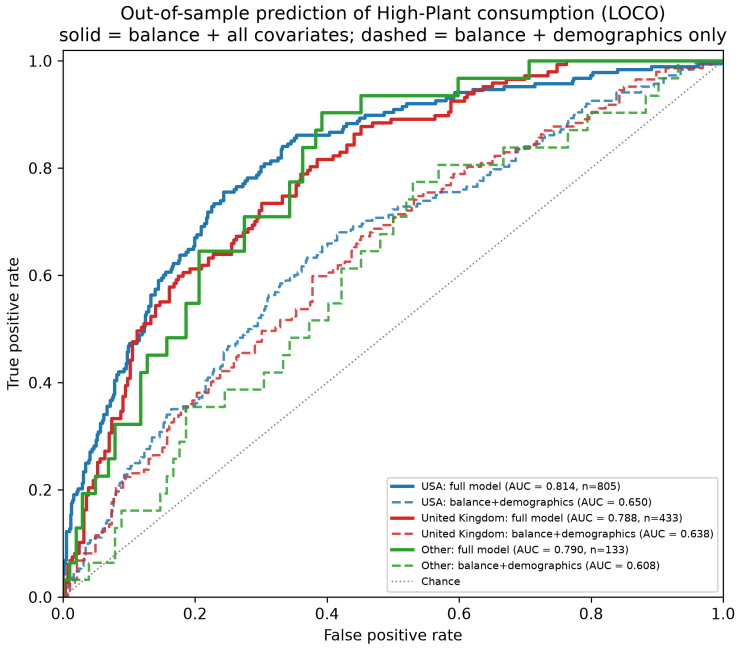


Figure 8: Out-of-sample prediction of High-Plant status in held-out countries. Solid curves (balance + all covariates) reach AUC 0.79–0.81, but dashed curves (balance + demographics, dietary self-report removed) fall to AUC 0.61–0.65. Because a covariates-only model already achieves AUC ≈ 0.82, the microbial balance adds only -0.019 AUC on average—its apparent predictive value is driven by dietary self-report, not by the microbiome.

3.6 Power and sample size for a redesigned trial

The gap between the in-sample and transportable effect sizes drives the trial arithmetic (Figure 9; Table 2). Powering optimistically on the in-sample $d = 0.602$, a parallel-group feeding study would need only ∼45 participants per group (90 total) for 80% power. But because the LOCO analysis shows this effect may not survive in a new population, the defensible choice is to power on the conservative, transportability-adjusted $d = 0.301$, which inflates the parallel-group requirement to 175 per group (350 total) at 80% power and 233 per group (466 total) at 90%. A within-subject crossover design, in which each participant serves as their own control and subject-specific microbiome composition—the dominant source of variance—is differenced out, is far more efficient: at a plausible between-period correlation of $r = 0.7$ and the same conservative effect, only ∼54 subjects achieve 80% power (72 for 90%), and even at a modest $r = 0.5$ only ∼89 subjects are required. The crossover thus attains the same power as a ∼350-participant parallel trial with roughly one-sixth the enrolment.

Table 2: Sample-size requirements for a future controlled-feeding study at two-sided $\alpha = 0.05$, comparing the optimistic in-sample effect ($d = 0.602$) with the conservative transportability-adjusted effect ($d = 0.301$). Crossover figures are numbers of subjects; parallel figures are per group.

Design	Effect size	80% power	90% power
Parallel (per group)	In-sample $d = 0.602$	45	59
Parallel (per group)	Conservative $d = 0.301$	**175**	233
Crossover, $r = 0.5$ (subjects)	Conservative $d = 0.301$	89	118
Crossover, $r = 0.7$ (subjects)	Conservative $d = 0.301$	**54**	72
Crossover, $r = 0.7$ (subjects)	In-sample $d = 0.602$	16	20

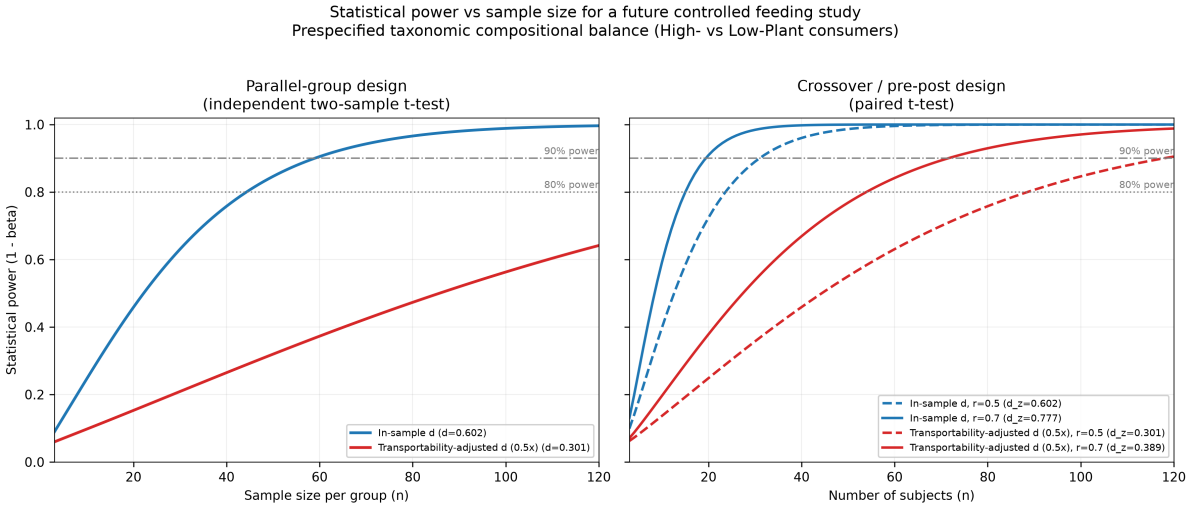


Figure 9: Statistical power versus sample size for a future controlled-feeding study on the prespecified taxonomic balance. Left: parallel-group design; the conservative effect (red) requires ∼175 per group for 80% power versus ∼45 for the optimistic in-sample effect (blue). Right: crossover design; within-subject differencing sharply reduces requirements, to ∼54 subjects at $r = 0.7$ under the conservative effect.

3.7 Recommendation and prespecified endpoint

Synthesizing the evidence chain—severe, geographically structured missingness (Step 2); a significant but negligible community-wide effect (Step 3); a strong, BMI-robust, internally cross-validated in-sample balance (Step 4); and a decisive failure to transport that balance across geography (Step 5)—the formal decision is **REDESIGN**, not *fund* and not *do-not-fund* (Box 1). A straight *fund* decision would assume the in-sample $d \approx 0.60$ transports to a new population, which the LOCO analysis directly falsifies. A *do-not-fund* decision would ignore that the endpoint is genuinely, highly significant in-sample and robust to covariate and BMI adjustment; the obstacles—geographic heterogeneity and reliance on error-prone dietary self-report—are exactly the confounds a controlled feeding study with a within-subject design is built to eliminate.

Box 1. Decision: REDESIGN a controlled-feeding study.
One-line rationale. The endpoint is real and highly significant in-sample ($p = 8.3 \times 10^{-12}$, robust to BMI) but does not transport across geography (held-out $\beta \approx 0$; no marginal AUC

over dietary self-report); a naive parallel-group trial powered on the raw association would likely fail, so fund only a redesigned trial.

Redesign requirements.

- Use a **crossover (within-subject)** design so each participant is their own baseline, controlling subject-specific microbiome composition and sharply reducing required n.
- **Stratify/block by geography** (country/site) and model site as a fixed effect, because the effect did not transport across USA / UK / Other in LOCO.
- Power on the **conservative transportability-adjusted effect** ($0.5\times$ in-sample d), not the optimistic in-sample estimate ($\sim$54 subjects at $r = 0.7$).
- **Prespecify and register** the exact balance endpoint below to prevent re-selection/overfitting.
- **Control the diet directly** rather than relying on self-report, given the severe missingness in Step 2.
- Restrict inference to **taxonomic composition**; do not infer metabolic or functional pathways from 16S data.

Prespecified primary endpoint. Mean change from baseline to end-of-intervention in the log-ratio balance $\mathrm{B} = \overline{\ln x_{\mathrm{num}}} - \overline{\ln x_{\mathrm{den}}}$ (genus-level, pseudocount 0.5), *numerator* {*o__Clostridiales*, *f__S24-7*, *g__Coprobacillus*, *g__Clostridium*} over *denominator* {*g__Blautia*, *g__Dorea*, *g__Anaerotruncus*, *f__Desulfovibrionaceae*}, compared between high- and low-plant controlled diets. Measurement platform: 16S rRNA amplicon sequencing (Greengenes genus level). No functional/metabolic activity is inferred.

4 Discussion

4.1 Principal findings

Re-examining the influential "30 plants per week" heuristic with compositional statistics on the fixed 2018 AGP fecal dataset yields a nuanced verdict. The plant-diversity contrast is reliably *detectable*: it survives adjustment for age, sex, BMI, geography, antibiotics and three other dietary items, and it survives permutation that holds geography fixed. Yet at the whole-community level the effect is vanishingly small ($R^2 \approx 0.1\%$ of Aitchison beta-diversity), and when concentrated into an interpretable, internally cross-validated eight-genus balance with a respectable in-sample effect ($d = 0.60$, $p = 8.3 \times 10^{-12}$), it fails the single most important test for a would-be biomarker: it does not transport to a held-out geographic cohort. The held-out coefficient is statistically indistinguishable from zero in all three folds, and the microbial balance adds essentially nothing to out-of-sample prediction once dietary self-report is accounted for. The signal is therefore best understood as a genuine but geographically contingent correlate of a self-reported exposure, not as a portable, mechanistic microbial fingerprint of plant diversity.

4.2 Why a strong in-sample effect fails to transport

The dissociation between a highly significant in-sample association and a null out-of-sample effect is a textbook manifestation of the generalization problem that now pervades microbiome research

[23, 26, 30]. Several mutually reinforcing mechanisms are at play here. First, geography is among the strongest known determinants of gut community structure [26–28], and the AGP cohort is overwhelmingly USA-based with a substantial UK stratum; a balance tuned on one geographic mixture need not hold in another. Second, the exposure is self-reported through a single coarse questionnaire item, and self-report is a noisy, culturally patterned proxy for actual plant intake—so part of the "plant-diversity" signal is inevitably confounded with country-level differences in diet reporting, cuisine and lifestyle. The attribution analysis makes this concrete: the out-of-sample AUC is driven by dietary self-report covariates, not by the microbiome. Third, unmeasured technical structure is a live concern. The AGP mapping file lacks any sequencing-plate identifier, so batch effects—demonstrated repeatedly to rival biological signal [24, 25]—cannot be modelled, and geography is the only available proxy for the correlated bundle of processing batch and population. That the taxa themselves reshuffle between folds (mean Jaccard 0.35) is exactly what one expects when a supervised model latches onto locally predictive but non-portable features. None of these mechanisms implies the underlying biology is absent; they imply that observational AGP data cannot isolate it.

4.3 Rationale for a redesigned, within-subject trial

The remedy is a study design that removes precisely these confounds. A controlled-feeding study replaces error-prone self-report with a directly administered, quantified diet, addressing the exposure-measurement problem that Step 2 exposed. A *within-subject crossover* additionally differences out each participant's idiosyncratic baseline community—the dominant axis of variation in every human microbiome survey [4, 44]—so that the plant-diet contrast is estimated against a within-person, rather than between-person, background. This is not merely a power optimization, though the power gain is dramatic (roughly 54 versus 350 participants; Table 2); it is a bias-reduction strategy, because the subject-specific and much of the geographic confounding that sank the LOCO analysis are held constant by construction. The empirical basis for expecting a real, measurable within-person response is strong: controlled and longitudinal feeding studies show that diet reconfigures the community rapidly and reproducibly [12, 44, 45], and fiber/plant substrates are the canonical drivers of fermentative taxa, with sustained fiber deprivation capable of depleting them [8, 10, 15, 46]. Powering on the conservative, transportability-adjusted effect—rather than the optimistic in-sample d—guards against designing a well-powered trial for an effect that may shrink in a fresh population, and geographic stratification with site as a fixed effect protects against the residual heterogeneity that a single-site or pooled analysis would reintroduce.

4.4 A single, prespecified, taxonomy-only endpoint

Registering one primary endpoint—the exact eight-genus balance identified here—is essential to prevent the re-selection and over-fitting that produce non-reproducible microbiome signatures [20, 29]. We deliberately frame this endpoint in strictly taxonomic terms. Although the numerator and denominator taxa include lineages often discussed in the context of fiber fermentation and bile metabolism, 16S amplicon data measure marker-gene abundance, not gene content, expression or metabolite flux, and tools that predict function from 16S are explicitly hypothesis-generating rather than confirmatory [21, 31, 32]. We therefore make no claim about short-chain fatty acids, enzymatic capacity or any metabolic pathway; the endpoint is a compositional log-ratio, and its interpretation is confined to the relative abundances of the named taxa. A future trial wishing to make functional claims would need shotgun metagenomics, metatranscriptomics or direct metabolomics as separate,

prespecified secondary measures.

4.5 Strengths and limitations

The analysis has several strengths: it uses the complete fixed AGP fecal cohort rather than the original small subgroup; it respects the compositional nature of the data throughout; it holds geography fixed both in permutation testing and in an explicit LOCO framework; and it separates statistical significance from effect size and from transportability, three properties that are too often conflated. Limitations temper the conclusions. The exposure and covariates are self-reported and, as shown, heavily missing, so complete-case models are susceptible to the geographically structured selection bias we document; although the distributions of complete and incomplete cases overlapped, formal multiple-imputation or inverse-probability weighting would be a valuable sensitivity extension. The absence of a sequencing-plate variable is a genuine gap that leaves batch effects only partially controlled through geography. The cohort's strong USA/UK skew limits the granularity of the LOCO test to three pooled strata, and the "Other" held-out fold is small ($n_{\text{valid}} = 133$), widening its confidence interval. Finally, because AGP is cross-sectional, no causal or temporal claim can be made—which is itself the core argument for the interventional redesign we recommend.

4.6 Conclusion

Within the American Gut Project, eating more than 30 versus 10 or fewer plant types per week is statistically associated with gut microbial ecology after broad covariate adjustment, and this association can be distilled into a strong, internally reproducible taxonomic balance. But the association is negligibly small community-wide, is carried out-of-sample by dietary self-report rather than by the microbiome, and—most tellingly—does not transport across geographic cohorts. These properties place the question squarely in "redesign" territory: the biology plausibly exists but cannot be established, let alone measured for effect size, from confounded observational data. We recommend funding a geographically stratified, within-subject crossover feeding study of approximately 54 participants, powered on a conservative transportability-adjusted effect, with a single prespecified taxonomic balance as its primary endpoint and with all functional interpretation deferred to purpose-built metagenomic or metabolomic measurements.

Data and code availability

The primary data are the fixed public AGP artifacts on Figshare: the fecal sOTU table (DOI 10.6084/m9.figshare.6137192) and the mapping file (DOI 10.6084/m9.figshare.6137315). All analysis scripts, JSON result logs and figures generated for this study are provided with the project.

Author contributions

The analysis pipeline, statistical modelling, figure generation and manuscript were produced as a single reproducible workflow.

Conflicts of interest

The authors declare no competing interests.

Ethics

This study is a secondary analysis of de-identified, publicly released American Gut Project data and did not involve new human-subjects data collection.

References

[1] Ron Sender, Shai Fuchs, and Ron Milo. Revised estimates for the number of human and bacteria cells in the body. *PLOS Biology*, 14(8):e1002533, 2016. doi: 10.1371/journal.pbio.1002533.

[2] Jack A. Gilbert, Martin J. Blaser, J. Gregory Caporaso, Janet K. Jansson, Susan V. Lynch, and Rob Knight. Current understanding of the human microbiome. *Nature Medicine*, 24(4): 392–400, 2018. doi: 10.1038/nm.4517.

[3] Jason Lloyd-Price, Galeb Abu-Ali, and Curtis Huttenhower. The healthy human microbiome. *Genome Medicine*, 8:51, 2016. doi: 10.1186/s13073-016-0307-y.

[4] Gwen Falony, Marie Joossens, Sara Vieira-Silva, Jun Wang, Youssef Darzi, Karoline Faust, Alexander Kurilshikov, Marc Jan Bonder, Mireia Valles-Colomer, Doris Vandeputte, et al. Population-level analysis of gut microbiome variation. *Science*, 352(6285):560–564, 2016. doi: 10.1126/science.aad3503.

[5] Alexandra Zhernakova, Alexander Kurilshikov, Marc Jan Bonder, Ettje F. Tigchelaar, Melanie Schirmer, Tommi Vatanen, Zlatan Mujagic, Arnau Vich Vila, Gwen Falony, Sara Vieira-Silva, et al. Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. *Science*, 352(6285):565–569, 2016. doi: 10.1126/science.aad3369.

[6] Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature*, 486(7402):207–214, 2012. doi: 10.1038/nature11234.

[7] Peter J. Turnbaugh, Micah Hamady, Tanya Yatsunenko, Brandi L. Cantarel, Alexis Duncan, Ruth E. Ley, Mitchell L. Sogin, William J. Jones, Bruce A. Roe, Jason P. Affourtit, et al. A core gut microbiome in obese and lean twins. *Nature*, 457(7228):480–484, 2009. doi: 10.1038/nature07540.

[8] Erica D. Sonnenburg and Justin L. Sonnenburg. Starving our microbial self: The deleterious consequences of a diet deficient in microbiota-accessible carbohydrates. *Cell Metabolism*, 20(5): 779–786, 2014. doi: 10.1016/j.cmet.2014.07.003.

[9] Ara Koh, Filipe De Vadder, Petia Kovatcheva-Datchary, and Fredrik Bäckhed. From dietary fiber to host physiology: Short-chain fatty acids as key bacterial metabolites. *Cell*, 165(6): 1332–1345, 2016. doi: 10.1016/j.cell.2016.05.041.

[10] Kassem Makki, Edward C. Deehan, Jens Walter, and Fredrik Bäckhed. The impact of dietary fiber on gut microbiota in host health and disease. *Cell Host & Microbe*, 23(6):705–715, 2018. doi: 10.1016/j.chom.2018.05.012.

[11] Gary D. Wu, Jun Chen, Christian Hoffmann, Kyle Bittinger, Ying-Yu Chen, Sue A. Keilbaugh, Meenakshi Bewtra, Dan Knights, William A. Walters, Rob Knight, et al. Linking long-term dietary patterns with gut microbial enterotypes. *Science*, 334(6052):105–108, 2011. doi: 10.1126/science.1208344.

[12] Lawrence A. David, Corinne F. Maurice, Rachel N. Carmody, David B. Gootenberg, Julie E. Button, Benjamin E. Wolfe, Alisha V. Ling, A. Sloan Devlin, Yug Varma, Michael A. Fischbach, et al. Diet rapidly and reproducibly alters the human gut microbiome. *Nature*, 505(7484): 559–563, 2014. doi: 10.1038/nature12820.

[13] Daniel McDonald, Embriette Hyde, Justine W. Debelius, James T. Morton, Antonio Gonzalez, Gail Ackermann, Alexey A. Aksenov, Bahar Behsaz, Caitriona Brennan, Yingfeng Chen, et al. American gut: An open platform for citizen science microbiome research. *mSystems*, 3(3): e00031–18, 2018. doi: 10.1128/mSystems.00031-18.

[14] Susan Mills, Catherine Stanton, Jonathan A. Lane, Graeme J. Smith, and R. Paul Ross. Precision nutrition and the microbiome, part i: Current state of the science. *Nutrients*, 11(4): 923, 2019. doi: 10.3390/nu11040923.

[15] Daniel So, Kevin Whelan, Megan Rossi, Mark Morrison, Gerald Holtmann, Jaimon T. Kelly, Erin R. Shanahan, Heidi M. Staudacher, and Katrina L. Campbell. Dietary fiber intervention on gut microbiota composition in healthy adults: A systematic review and meta-analysis. *The American Journal of Clinical Nutrition*, 107(6):965–983, 2018. doi: 10.1093/ajcn/nqy041.

[16] John Aitchison. The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(2):139–177, 1982. doi: 10.1111/j.2517-6161.1982.tb01195.x.

[17] Gregory B. Gloor, Jean M. Macklaim, Vera Pawlowsky-Glahn, and Juan J. Egozcue. Microbiome datasets are compositional: And this is not optional. *Frontiers in Microbiology*, 8:2224, 2017. doi: 10.3389/fmicb.2017.02224.

[18] Paul J. McMurdie and Susan Holmes. Waste not, want not: Why rarefying microbiome data is inadmissible. *PLOS Computational Biology*, 10(4):e1003531, 2014. doi: 10.1371/journal.pcbi.1003531.

[19] Juan José Egozcue, Vera Pawlowsky-Glahn, Glòria Mateu-Figueras, and Carles Barceló-Vidal. Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, 35 (3):279–300, 2003. doi: 10.1023/A:1023818214614.

[20] J. Rivera-Pinto, Juan J. Egozcue, Vera Pawlowsky-Glahn, R. Paredes, M. Noguera-Julian, and M. Luz Calle. Balances: A new perspective for microbiome analysis. *mSystems*, 3(4):e00053–18, 2018. doi: 10.1128/mSystems.00053-18.

[21] James T. Morton, Clarisse Marotz, Alex Washburne, Justin Silverman, Livia S. Zaramela, Anna Edlund, Karsten Zengler, and Rob Knight. Establishing microbial composition measurement standards with reference frames. *Nature Communications*, 10:2719, 2019. doi: 10.1038/s41467-019-10656-5.

[22] Justin D. Silverman, Alex D. Washburne, Sayan Mukherjee, and Lawrence A. David. A phylogenetic transform enhances analysis of compositional microbiota data. *eLife*, 6:e21887, 2017. doi: 10.7554/eLife.21887.

[23] Ivan Vujkovic-Cvijin, Jack Sklar, Lingjing Jiang, Loki Natarajan, Rob Knight, and Yasmine Belkaid. Host variables confound gut microbiota studies of human disease. *Nature*, 587(7834): 448–454, 2020. doi: 10.1038/s41586-020-2881-9.

[24] Rashmi Sinha, Galeb Abu-Ali, Emily Vogtmann, Anthony A. Fodor, Boyu Ren, Amnon Amir, Emma Schwager, Jonathan Crabtree, Siyuan Ma, Christian C. Abnet, et al. Assessment of variation in microbial community amplicon sequencing by the microbiome quality control (mbqc) project consortium. *Nature Biotechnology*, 35(11):1077–1086, 2017. doi: 10.1038/nbt.3981.

[25] Yiwen Wang and Kim-Anh Lê Cao. Managing batch effects in microbiome data. *Briefings in Bioinformatics*, 21(6):1954–1970, 2020. doi: 10.1093/bib/bbz105.

[26] Yan He, Wei Wu, Hui-Min Zheng, Pan Li, Daniel McDonald, Hua-Fang Sheng, Mu-Xuan Chen, Zhi-Hong Chen, Gui-Yuan Ji, Zhang-Dong-Xie Zheng, et al. Regional variation limits applications of healthy gut microbiome reference ranges and disease models. *Nature Medicine*, 24(10):1532–1535, 2018. doi: 10.1038/s41591-018-0164-x.

[27] Tanya Yatsunenko, Federico E. Rey, Mark J. Manary, Indi Trehan, Maria Gloria Dominguez-Bello, Monica Contreras, Magda Magris, Glida Hidalgo, Robert N. Baldassano, Andrey P. Anokhin, et al. Human gut microbiome viewed across age and geography. *Nature*, 486(7402): 222–227, 2012. doi: 10.1038/nature11053.

[28] Pajau Vangay, Abigail J. Johnson, Tonya L. Ward, Gabriel A. Al-Ghalith, Robin R. Shields-Cutler, Benjamin M. Hillmann, Sarah K. Lucas, Lalit K. Beura, Emily A. Thompson, Lisa M. Till, et al. Us immigration westernizes the human gut microbiome. *Cell*, 175(4):962–972.e10, 2018. doi: 10.1016/j.cell.2018.09.029.

[29] Chloe Mirzayi, Audrey Renson, Fatima Zohra, Shaimaa Elsafoury, Ludwig Geistlinger, Lora J. Kasselman, Kelly Eckenrode, Janneke van de Wijgert, Amy Loughman, Francine Z. Marques, et al. Reporting guidelines for human microbiome research: The storms checklist. *Nature Medicine*, 27(11):1885–1892, 2021. doi: 10.1038/s41591-021-01552-x.

[30] Marc A. Sze and Patrick D. Schloss. Looking for a signal in the noise: Revisiting obesity and the microbiome. *mBio*, 7(4):e01018–16, 2016. doi: 10.1128/mBio.01018-16.

[31] Morgan G. I. Langille, Jesse Zaneveld, J. Gregory Caporaso, Daniel McDonald, Dan Knights, Joshua A. Reyes, Jose C. Clemente, Deron E. Burkepile, Rebecca L. Vega Thurber, Rob Knight, Robert G. Beiko, and Curtis Huttenhower. Predictive functional profiling of microbial communities using 16s rrna marker gene sequences. *Nature Biotechnology*, 31(9):814–821, 2013. doi: 10.1038/nbt.2676.

[32] Gavin M. Douglas, Vincent J. Maffei, Jesse R. Zaneveld, Svetlana N. Yurgel, James R. Brown, Christopher M. Taylor, Curtis Huttenhower, and Morgan G. I. Langille. Picrust2 for prediction of metagenome functions. *Nature Biotechnology*, 38(6):685–688, 2020. doi: 10.1038/s41587-020-0548-6.

[33] Amnon Amir, Daniel McDonald, Jose A. Navas-Molina, Evguenia Kopylova, James T. Morton, Zhenjiang Zech Xu, Eric P. Kightley, Luke R. Thompson, Embriette R. Hyde, Antonio Gonzalez, and Rob Knight. Deblur rapidly resolves single-nucleotide community sequence patterns. *mSystems*, 2(2):e00191–16, 2017. doi: 10.1128/mSystems.00191-16.

[34] Benjamin J. Callahan, Paul J. McMurdie, Michael J. Rosen, Andrew W. Han, Amy Jo A. Johnson, and Susan P. Holmes. Dada2: High-resolution sample inference from illumina amplicon data. *Nature Methods*, 13(7):581–583, 2016. doi: 10.1038/nmeth.3869.

[35] Evan Bolyen, Jai Ram Rideout, Matthew R. Dillon, Nicholas A. Bokulich, Christian C. Abnet, Gabriel A. Al-Ghalith, Harriet Alexander, Eric J. Alm, Manimozhiyan Arumugam, Francesco Asnicar, et al. Reproducible, interactive, scalable and extensible microbiome data science using qiime 2. *Nature Biotechnology*, 37(8):852–857, 2019. doi: 10.1038/s41587-019-0209-9.

[36] Todd Z. DeSantis, Philip Hugenholtz, Neils Larsen, Mark Rojas, Eoin L. Brodie, Keith Keller, Thomas Huber, Daniel Dalevi, Ping Hu, and Gary L. Andersen. Greengenes, a chimera-checked 16s rrna gene database and workbench compatible with arb. *Applied and Environmental Microbiology*, 72(7):5069–5072, 2006. doi: 10.1128/AEM.03006-05.

[37] Daniel McDonald, Yoshiki Vazquez-Baeza, David Koslicki, Jamie McClelland, Nicolai Reeve, Zhenjiang Xu, Antonio Gonzalez, and Rob Knight. Striped unifrac: Enabling microbiome analysis at unprecedented scale. *Nature Methods*, 15(11):847–848, 2018. doi: 10.1038/s41592-018-0187-8.

[38] Marti J. Anderson. A new method for non-parametric multivariate analysis of variance. *Austral Ecology*, 26(1):32–46, 2001. doi: 10.1111/j.1442-9993.2001.tb00081.x.

[39] Pierre Legendre and Marti J. Anderson. Distance-based redundancy analysis: Testing multispecies responses in multifactorial ecological experiments. *Ecological Monographs*, 69(1):1–24, 1999. doi: 10.1890/0012-9615(1999)069[0001:DBRATM]2.0.CO;2.

[40] Marti J. Anderson and Daniel C. I. Walsh. Permanova, anosim, and the mantel test in the face of heterogeneous dispersions: What null hypothesis are you testing? *Ecological Monographs*, 83 (4):557–574, 2013. doi: 10.1890/12-2010.1.

[41] Skipper Seabold and Josef Perktold. Statsmodels: Econometric and statistical modeling with python. *Proceedings of the 9th Python in Science Conference*, pages 92–96, 2010. doi: 10.25080/Majora-92bf1922-011.

[42] Jacob Cohen. A power primer. *Psychological Bulletin*, 112(1):155–159, 1992. doi: 10.1037/0033-2909.112.1.155.

[43] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. doi: 10.5555/1953048.2078195.

[44] Abigail J. Johnson, Pajau Vangay, Gabriel A. Al-Ghalith, Benjamin M. Hillmann, Tonya L. Ward, Robin R. Shields-Cutler, Austin D. Kim, Anna K. Shmagel, Abdul N. Syed, Jens Walter, et al. Daily sampling reveals personalized diet-microbiome associations in humans. *Cell Host & Microbe*, 25(6):789–802.e5, 2019. doi: 10.1016/j.chom.2019.05.005.

[45] Hannah C. Wastyk, Gabriela K. Fragiadakis, Dalia Perelman, Devangi Dahan, Bryan D. Merrill, Feiqiao B. Yu, Madeline Topf, Carlos G. Gonzalez, William Van Treuren, Shuo Han, et al. Gut-microbiota-targeted diets modulate human immune status. *Cell*, 184(16):4137–4153.e14, 2021. doi: 10.1016/j.cell.2021.06.019.

[46] Erica D. Sonnenburg, Samuel A. Smits, Mikhail Tikhonov, Steven K. Higginbottom, Ned S. Wingreen, and Justin L. Sonnenburg. Diet-induced extinctions in the gut microbiota compound over generations. *Nature*, 529(7585):212–215, 2016. doi: 10.1038/nature16504.