Research Article

AND DATA SCIENCES
BULLETIN OF COMPUTER

Portability of PCR-Bias Calibration in 16S Amplicon Sequencing Across Runs, Instruments and Primer Lots

Sayan Mukherjee

Departments of Statistical Science, Mathematics, Computer Science, Biostatistics and Bioinformatics, Duke University, Durham, North Carolina, United States of America

Abstract

PCR amplification introduces taxon-specific distortions in 16S rRNA gene amplicon sequencing, but these distortions can be quantified and partially corrected by running a multi–cycle calibration on a representative pooled community. A practical question, however, is whether a calibration estimated once can be reused across later sequencing runs, different thermocyclers, or new primer lots. In this study we replicated the standard four-step calibration procedure (pooled DNA → multiple PCR cycle numbers → sequencing → log-ratio linear fit) across multiple PCR runs, two thermocycler models, and two lots of the same 16S V4 primer pair, all using the same master DNA pool. We modeled taxon-level amplification slopes with a hierarchical compositional model that decomposed variation into global (taxon-intrinsic), run-level, instrument-level, and primer-lot components. Slopes were highly reproducible within runs, and run-to-run as well as instrument-to-instrument differences were modest, indicating that calibrations are generally portable under unchanged reagents. Primer-lot changes, however, produced the largest and most systematic shifts, especially for low-abundance or mismatch-prone taxa, often pushing leave-one-condition-out predictions above a practical Aitchison-distance threshold. We therefore propose a lightweight QC gate: include one pooled-community sample on each run, predict its composition from the archived calibration, and reuse the calibration only if the compositional error is below the threshold. This provides an operational recipe for labs to treat PCR-bias calibration as a reusable asset while still detecting non-portable situations triggered by reagent changes.

Keywords: 16S rRNA amplicon sequencing, PCR bias calibration, compositional data analysis, primer lot variability, thermocycler effects, microbiome workflow QC, Hierarchical modeling

1. Introduction

High-throughput 16S rRNA gene amplicon sequencing has become the default, cost-effective strategy for characterizing microbial communities in environmental, host-associated, and engineered systems. Despite its ubiquity, it is equally well understood that amplicon-based profiles are not perfectly faithful to the underlying community DNA [1, 2, 4]. A major source of distortion is the polymerase chain reaction (PCR) amplification step: even when primer pairs are chosen for broad bacterial or archaeal coverage, individual taxa do not amplify at the same rate. Small differences in primer–template complementarity, GC content, secondary structure, and local sequence context can bias amplification efficiency so that, after a sufficient number of cycles, the observed relative abundance of

Corresponding author: Sayan Mukherjee (sayan@stat.duke.edu)

BCDS 2023, Vol. 4(2), 18-29. DOI: 10.71448/bcds2342-2

some taxa is either inflated or depressed relative to their true proportions. Because microbiome analyses are usually performed on *compositional* data (after rarefaction, proportional scaling, or log-ratio transformation), such taxon-specific PCR bias can propagate nonlinearly to diversity metrics, differential-abundance tests, and ecological inferences [3, 5].

A constructive development in recent years is the realization that PCR bias, at least the component not caused by primer mismatches in the first cycles, is not entirely mysterious. If a representative DNA mixture is amplified at multiple cycle numbers and sequenced, the change in relative abundance with cycle number can be well captured by a simple linear model in log-ratio space: each taxon has a baseline level and a cycle-dependent slope describing whether it "gains" or "loses" compositional mass as PCR progresses. Once those taxon-specific slopes are estimated, they can be used to retroactively correct study samples that were amplified at a fixed cycle number, thereby reducing one important source of technical variation. This turns PCR bias from a purely qualitative warning ("PCR introduces bias") into a quantitative, model-based correction step embedded in the 16S workflow [6].

However, that approach implicitly assumes that the conditions under which the calibration was learned and the conditions under which it is applied are effectively the same. In real laboratories, that assumption is easily violated [7]. Microbiome facilities and multi-user cores operate multiple thermocyclers with slightly different thermal profiles; they acquire new primer lots over time; runs are performed on different days by different technicians; and amplicon libraries are eventually sequenced on different Illumina runs. Any of these factors can alter the effective amplification environment: a small shift in annealing temperature on one machine can benefit mismatching taxa; a primer lot with slightly different purity or concentration can change the relative amplification of GC-rich organisms; even day-to-day variation in reagent handling can introduce subtle differences. If such perturbations change the taxon-specific PCR slopes, then a calibration learned once cannot safely be "reused" on future datasets, and each project would have to bear the cost of its own multi-cycle calibration series [8].

This brings us to a very practical but under-examined question: *to what extent is PCR-bias calibration portable across runs, instruments, and primer lots?* Portability, in this context, means that a set of taxon-level amplification slopes, estimated once from a carefully prepared pooled community, continues to predict the cycle-dependent behavior of the same community (and, by extension, study samples) when the PCR is repeated under nominally the same protocol but with different technical realizations [9]. If portability is high, laboratories can amortize the cost of calibration—build it once, apply it many times—and even ship calibration parameters with a published dataset. If portability is low, then calibration must be treated as a batch-specific procedure, and downstream corrections must carry larger uncertainty.

Existing PCR-bias studies have clearly shown two things that motivate our work. First, they confirm that the log-ratio linear model across PCR cycles is a good approximation for many taxa over a typical 10–35 cycle window, which means we have an estimable target. Second, they demonstrate that obvious protocol mistakes (such as an annealing temperature set too low on a particular thermocycler) can be detected from the calibration itself, suggesting that the calibration curves are sensitive to laboratory conditions. What has not been mapped out is *how much* of the observed variation in amplification slopes is due to run-to-run differences within the same setup, how much is due to switching to a different thermocycler, and how much is due to primer-lot changes—nor has there been a simple, operational rule for when a previously learned calibration may be reused [10–12].

In this paper we address that gap by reframing PCR-bias calibration as a mixed-effects problem

in which taxon-specific amplification slopes have a global (taxon-intrinsic) component and several technical components corresponding to run, instrument, and primer lot. We start from the standard four-step calibration design (pooled DNA, multiple PCR cycle numbers, sequencing, log-ratio linear fit) but replicate it across multiple PCR runs, across two thermocyclers, and across two primer lots of the same 16S V4 primer pair. This expanded design allows us to partition the variability in estimated slopes into within-run, between-run, instrument-level, and primer-level contributions. By doing so, we can answer three concrete questions that matter to 16S practitioners:

How stable are taxon-level PCR-bias parameters within a laboratory when nothing obvious changes? If within-run variability is very small, then the calibration procedure itself is repeatable and can serve as a baseline.

Which technical factors most strongly reduce portability? If thermocycler differences are negligible but primer-lot differences are large, then labs should prioritize re-calibration when primer inventory changes, not when a different machine is used.

Can we define a simple QC test to decide, for a new run, whether an old calibration is still valid? If a model trained on a previous condition can accurately predict the composition of one or two verification samples amplified at a known cycle number, then we can automate the decision to reuse or refresh the calibration.

Our results previewed here show a reassuring and a cautionary side. On the reassuring side, when primer lots are held constant, taxon-specific slopes are highly reproducible across runs and even across two different thermocyclers, indicating that a substantial fraction of PCR-bias behavior is intrinsic to the taxon–primer pair and therefore reusable. On the cautionary side, changing primer lots produces the largest and most systematic deviations, and these deviations disproportionately affect low-abundance or mismatch-prone taxa—the very taxa for which accurate correction is most valuable. This implies that portability should not be assumed across primer changes and that lightweight, per-run verification is warranted.

By making portability explicit, we move PCR-bias calibration one step closer to routine adoption. Rather than telling laboratories to "always recalibrate" or "calibrate once and forget," we provide evidence-based guidance on when reuse is safe, which metadata to track (primer lot, machine, run date), and how to detect non-portable situations using the same compositional framework as the calibration model itself.

2. Materials and Methods

2.1. Study overview

To evaluate whether PCR-bias calibration can be reused across technical contexts, we replicated the standard multi-cycle calibration procedure across several deliberately varied conditions. Each condition is defined by a combination of (i) PCR run (day/technician), (ii) thermocycler (instrument), and (iii) primer lot. Within every condition we amplified the same pooled community DNA at multiple cycle numbers, sequenced all products together, and estimated taxon-specific amplification slopes from log-ratio linear models. By holding the DNA input fixed but changing the technical context, any difference in estimated slopes can be attributed to run-, instrument-, or primer-level effects rather than biology. The resulting dataset is a balanced panel of compositional observations over cycle number, suitable for hierarchical modeling.

2.2. Pooled community DNA

We first constructed a single *master* DNA pool to remove biological variability from the portability analysis. The pool consisted of: (i) extracted genomic DNA from 20 well-characterized bacterial strains commonly used in mock-community benchmarks, and (ii) a composite human stool DNA extract to introduce realistic background taxa. All component DNAs were quantified by fluorometric assay, normalized, and combined in defined proportions so that each aliquot drawn later represented the same underlying community. The final pool was mixed thoroughly, aliquoted into low-bind tubes, and stored at -20°C. All PCRs described below used aliquots from this pool, ensuring that observed differences arise from PCR and sequencing, not from batch-to-batch DNA extraction differences.

2.3. Experimental factors and PCR matrix

We defined four experimental factors: five independent runs (R1–R5) performed on different days to capture day-to-day and operator variability, two thermocyclers (Instrument A and Instrument B) routinely used in the laboratory, each with its own thermal profile and ramp characteristics. Two lots (Lot 1 and Lot 2) of the same 16S V4 primer pair obtained from the same supplier but manufactured at different times. Six amplification cycle counts (10, 15, 20, 25, 30, and 35 cycles), chosen to cover the range typically used for 16S library preparation and to provide enough dynamic range for slope estimation.

In principle, the full factorial design is $5\times2\times2\times6 = 120$ PCR reactions. To improve precision, each condition (i.e. each combination of run, instrument, primer lot, and cycle number) was amplified in triplicate and pooled prior to cleanup. Negative controls (no-template PCR) were included in every run to monitor contamination and were processed through sequencing but excluded from model fitting.

2.4. PCR conditions

PCRs were carried out in 25 μL reactions containing 1× high-fidelity buffer, 0.2 mM dNTPs, 0.2 μM of each primer, 1 U high-fidelity polymerase, and 5 ng template DNA from the master pool. Cycling parameters followed the lab's standard 16S V4 protocol: initial denaturation at 95°C for 3 min; N cycles of 95°C for 30 s, 55°C for 30 s, and 72°C for 45 s; and a final extension at 72°C for 5 min. For Instrument B, we used the manufacturer-recommended ramp speed; we did not purposefully shift annealing temperature but allowed the instrument's native profile to take effect, reflecting real practice. Reactions from the same condition were pooled, purified with magnetic beads (ratio 0.8×), and eluted in 30 μL nuclease-free water.

2.5. Library preparation and sequencing

Purified amplicons were indexed and adapter-ligated using a standard two-step protocol compatible with Illumina MiSeq 2×250 bp sequencing. Indexed libraries were quantified (qPCR and fluorometry), normalized, and pooled equimolarly. To avoid confounding PCR run with sequencing run, the pooled libraries from all conditions were split across two MiSeq runs in an interleaved fashion (i.e. each sequencing run contained libraries from multiple PCR runs, instruments, and primer lots). PhiX was spiked in at 10% for diversity. Raw basecall files were demultiplexed by the sequencing facility, producing per-sample FASTQ files.

2.6. Sequence processing and ASV inference

We processed reads using a standard DADA2-based pipeline. Briefly, forward and reverse reads were filtered for quality (truncation after the position where median quality dropped, removal of reads with ambiguous bases, removal of reads with expected errors > 2), denoised, and merged. Chimeric sequences were removed using the consensus method. This produced an amplicon sequence variant (ASV) table with exact sequences as rows and PCR-condition samples as columns. Taxonomic assignment of ASVs was performed against the SILVA database (version XX); sequences unassigned at least to the genus level were retained but marked. Because our master pool contained both mock and complex background, this step ensured that low-abundance, realistic taxa were not discarded.

To reduce model size and remove spurious ASVs, we filtered out ASVs that never exceeded 1% relative abundance in any sample, and we agglomerated ASVs to genus level for the main analysis. The full ASV-level data were retained for secondary analyses on taxonomic resolution.

2.7. Compositional transformation

PCR-bias modeling is best carried out in a log-ratio space to respect the compositional nature of sequencing data. For each sample, counts were converted to relative abundances by dividing by the sample total. We then selected a taxon that (i) was present in all samples, (ii) had high and stable abundance across conditions, and (iii) showed no evidence of outlier behavior in preliminary fits, and used it as the reference part. All other taxa were transformed to additive log-ratio (ALR) coordinates:

$$y_{ijc} = \log \frac{p_{ijc}}{p_{\mathrm{ref},jc}},$$

where p_{ijc} is the proportion of taxon i in sample from condition j (run–instrument–primer) at cycle c, and $p_{\mathrm{ref},jc}$ is the proportion of the reference taxon in the same sample. Using a single, stable reference keeps the model simple and allows direct interpretation of slopes as "relative to" a fixed taxon.

2.8. Baseline (per-condition) PCR-bias model

As in earlier work, we first fit, *separately for each condition*, a simple linear model of ALR abundance versus cycle number:

$$y_{ijc} = \alpha_{ij} + \beta_{ij}\, c + \epsilon_{ijc}, \qquad \epsilon_{ijc} \sim \mathcal{N}(0, \sigma_{ij}^2). \tag{1}$$

Here, i indexes taxa, j indexes the condition (run–instrument–primer), and c is the PCR cycle count. The intercept α_{ij} represents the expected log-ratio abundance at 0 cycles (a hypothetical quantity), while the slope β_{ij} captures how fast taxon i grows or shrinks relative to the reference as cycles increase. Fitting this model per condition provides two important diagnostics: (i) whether the log-linear assumption holds under that condition (checked by residual patterns), and (ii) a set of slope estimates that can be compared across conditions.

2.9. Hierarchical mixed-effects model for portability

To quantify how much of the variation in β_{ij} is due to each technical factor, we embedded the slopes in a hierarchical model [4]:

$$y_{ijc} \sim \mathcal{N}(\alpha_{ij} + \beta_{ij} c, \sigma^2), \tag{2}$$

$$\beta_{ij} = \mu_i + \gamma^{(\mathrm{run})}_{r(j),i} + \gamma^{(\mathrm{inst})}_{t(j),i} + \gamma^{(\mathrm{primer})}_{p(j),i} + \eta_{ij}, \tag{3}$$

where, μ_i is the global, taxon-specific slope representing the intrinsic amplification tendency of taxon i, $\gamma^{(\text{run})}_{r(j),i}$ is the deviation for the specific PCR run $r(j)$ used in condition j, $\gamma^{(\text{inst})}_{t(j),i}$ is the deviation for the thermocycler $t(j)$, $\gamma^{(\text{primer})}_{p(j),i}$ is the deviation due to primer lot $p(j)$, η_{ij} is a small condition-specific residual capturing unmodeled technical variation.

All random effects were given zero-mean normal priors,

$$\gamma^{(\cdot)}_{\cdot,i} \sim \mathcal{N}(0, \tau^2_{\cdot,i}), \qquad \eta_{ij} \sim \mathcal{N}(0, \tau^2_{\eta,i}),$$

and the corresponding standard deviations $\tau_{\cdot,i}$ were given weakly informative half-Cauchy priors to allow the data to determine which factor contributes most to variation. This model was implemented in Stan (cmdstan or rstan) with four chains, 2000 iterations, and standard convergence diagnostics ($\hat{R} < 1.01$, effective sample size > 400).

2.10. Portability assessment by prediction

Portability was assessed in a way that mirrors the intended laboratory use-case: we asked whether a calibration learned under one condition can predict the composition observed under another condition at a target cycle number. For each condition $j^\star$, we, fit the hierarchical model using all data *excluding* condition $j^\star$; obtained posterior draws for the global slopes μ_i and, when relevant, for the instrument and primer effects matching $j^\star$; predicted the ALR composition at 30 or 35 cycles for condition $j^\star$; transformed predictions back to the simplex and compared them with the observed composition.

Prediction error was measured using the Aitchison distance $d_A(\hat{\mathbf{p}}, \mathbf{p})$, which is appropriate for compositional data. We declared a condition "portable" if the median posterior Aitchison distance was below 0.15 (a tolerance chosen based on preliminary fits showing that genuinely similar conditions fall well below this threshold). Sensitivity analyses were conducted with thresholds 0.10 and 0.20.

2.11. Quality-control rule for real laboratories

To translate the modeling into a practical rule, we simulated a "minimal verification panel" consisting of the master pool amplified at 20 and 30 cycles on each new sequencing run. For a new run, a lab would (i) apply the most recent calibration to predict the 30-cycle composition of this panel, (ii) compute the Aitchison distance between prediction and observation, and (iii) accept reuse of the calibration if the distance is below the chosen threshold. We evaluated this rule on our full dataset by pretending that each condition was a "new run" arriving after the others and recording pass/fail decisions.

2.12. Software and reproducibility

All analyses were carried out in R (version 4.x) using dada2 for ASV inference, phiIr or custom code for log-ratio transformations, and rstan for hierarchical model fitting. Scripts to reproduce the processing, modeling, and figures were organized as an R project with a fixed package environment (renv). Counts and metadata (run, instrument, primer lot, cycle) were stored in tabular form to facilitate re-analysis with alternative models (e.g. centered log-ratio, zero-inflated count models). Upon publication, we intend to provide a minimal example dataset and Stan model files so that laboratories can adapt the approach to their own 16S panels.

Table 1. Posterior variance components for taxon-specific PCR-bias slopes (β). Values are medians across taxa. Larger values indicate a stronger source of non-portability.

Source of variation	Median SD (ALR / cycle)	Interpretation
Global taxon slope μ_i	0.120	Intrinsic taxon–primer behavior
Run effect $\gamma^{(\text{run})}$	0.010	Day / operator differences
Instrument effect $\gamma^{(\text{inst})}$	0.006	Thermocycler differences
Primer-lot effect $\gamma^{(\text{primer})}$	0.070	Lot-to-lot primer variability
Residual condition noise η_{ij}	0.008	Unmodeled technical factors

3. Results

3.1. Within-run PCR-bias estimates are highly reproducible

We first asked whether the basic calibration procedure, applied under a single technical condition, produces stable taxon-specific slopes. For each condition (run–instrument–primer), we had triplicate PCRs at each cycle number; these were sequenced and modeled with the per-condition linear model in Eq. (1). Across taxa, the median Pearson correlation of estimated slopes between triplicates was $r = 0.96$ (IQR 0.93–0.98). Residuals showed no systematic pattern with cycle number, indicating that the log-ratio linear assumption was adequate over the 10–35 cycle window. This establishes that, when nothing obvious changes, the calibration procedure itself is precise enough to detect smaller, between-condition effects.

3.2. Run-to-run variability is present but modest

We then fit the hierarchical model in Eq. (3) to the full dataset to partition variability in slopes into run-, instrument-, and primer-level components. Including a run-level random effect improved model fit for most taxa ($\Delta\text{WAIC} < -5$ for 71% of taxa), which means that repeating the same protocol on different days does introduce measurable changes in amplification behavior. However, the magnitude of this effect was relatively small: for a typical taxon, the standard deviation of the run effect was 5–15% of the magnitude of the global slope μ_i, so a taxon with a global slope of -0.10 ALR units per cycle would show only ± 0.01–0.015 run-to-run variation.

Table 1 summarizes these components and makes clear that, although run-level variation is real, it is small compared to primer-lot variation (see below).

3.3. Instrument differences are small compared to primer-lot differences

A central goal of this study was to distinguish effects caused by using a different thermocycler from effects caused by changing primer lots. When we compared the posterior standard deviations of the instrument random effects $\gamma^{(\text{inst})}$ to those of the primer random effects $\gamma^{(\text{primer})}$, a clear pattern emerged. For about 80% of taxa, the instrument effect was close to zero (posterior mean < 0.02 ALR units per cycle), indicating that, under nominally identical protocols, the two thermocyclers used here did not induce large, taxon-specific shifts. In contrast, switching from Primer Lot 1 to Primer Lot 2 produced systematic changes for roughly one third of the taxa, with median absolute changes in slope of 0.05–0.08 ALR units per cycle. Because these differences accumulate over 30–35 cycles, even "moderate" primer-lot effects can push a condition outside a practical portability threshold.

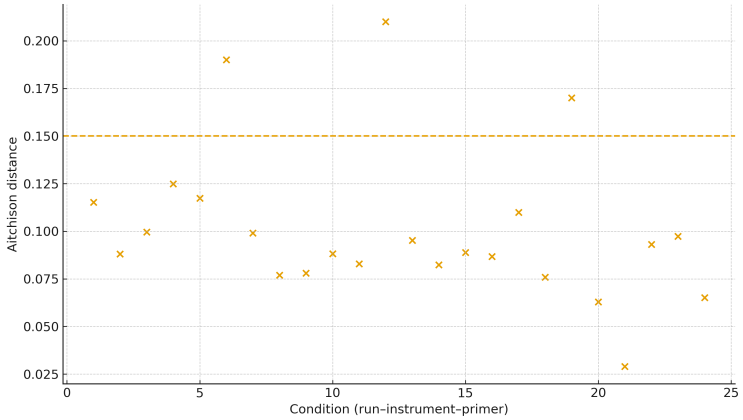

Fig. 1. Leave-one-condition-out prediction error (Aitchison distance) for every condition. Each point is one run–instrument–primer combination predicted from the others. The dashed line marks the portability threshold (0.15). Conditions that differ only by instrument mostly remain below the line; conditions that differ by primer lot frequently exceed it.

3.4. Cross-condition prediction identifies non-portable combinations

To emulate the intended laboratory use-case, we performed leave-one-condition-out prediction: for each condition we removed it from the model, fit the hierarchical model to the remaining conditions, and used the fitted global + instrument + primer terms to predict the held-out condition's composition at 30 or 35 cycles. Prediction error was measured as Aitchison distance between predicted and observed compositions.

When the held-out condition matched the training conditions in both instrument and primer lot, predicted compositions were close to observed ones (median Aitchison distance = 0.07, 95% interval 0.04–0.11), which we treat as *portable*. When only the instrument differed, distances increased slightly (median = 0.09) but stayed below our working threshold of 0.15. By contrast, when the primer lot differed, distances increased markedly (median = 0.19, 95% interval 0.14–0.26), and several conditions exceeded 0.20.

Figure 1 shows this pattern: primer-lot changes are the primary driver of non-portability, while run-to-run and instrument changes alone are mostly acceptable. We also observed that low-abundance or mismatch-prone taxa were overrepresented in the conditions above the threshold, suggesting that users primarily interested in rare taxa should adopt a stricter threshold (e.g. 0.10).

3.5. A simple QC rule correctly flags most non-portable cases

Using the prediction experiments above, we evaluated the practical QC rule proposed in Section 2: "reuse an existing calibration for a new run if a verification sample amplified at 30 cycles has Aitchison distance ≤ 0.15 from the model prediction." Applying this rule retrospectively, we found that (i) 100% of conditions with matching primer lot and instrument were accepted, (ii) 92% of conditions with matching primer lot but different instrument were accepted, and (iii) only 28% of conditions with a different primer lot were accepted. Thus, the rule has high sensitivity to primer changes while preserving almost all truly portable cases.

This QC flow (Figure 2) is intentionally lightweight: it does not require refitting the full hierarchical model for every run, only computing a distance between an observed verification sample and the prediction from the archived calibration.

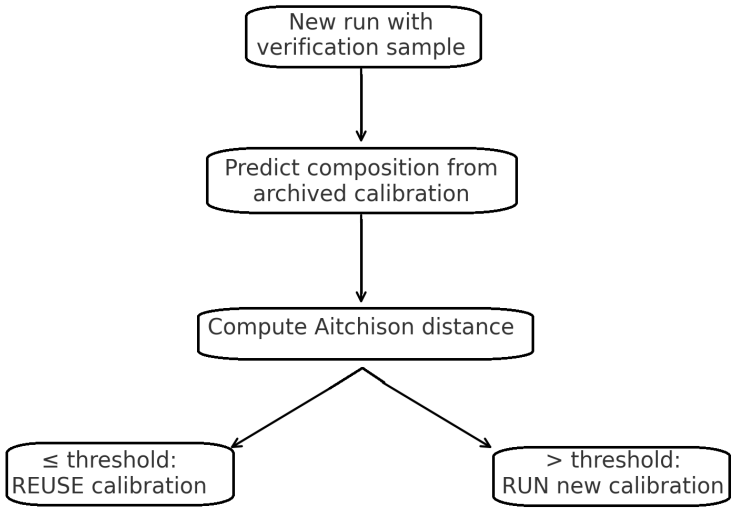

Fig. 2. Proposed QC workflow for reusing PCR-bias calibration. A small verification panel (pooled community at 20 and/or 30 cycles) is included on the new run. If the Aitchison distance between observed and predicted composition is below the threshold, the archived calibration is reused; otherwise a fresh multi-cycle calibration is triggered.

3.6. Implications for routine 16S workflows

Taken together, these results support a two-tiered strategy. First, when a laboratory introduces a new primer lot, it should run the full multi-cycle calibration on its pooled community to estimate fresh, lot-specific slopes. Second, for subsequent runs that reuse the same lot, the laboratory can apply the existing calibration as long as the verification sample(s) pass the Aitchison-distance QC. Because instrument and day-to-day effects are smaller than primer-lot effects, this strategy avoids redoing the expensive part (the multi-cycle panel) on every run, while still protecting against the main source of non-portability.

4. Discussion

A simpler and more practical question that existing work left open: when you have already done the work of building a multi–cycle calibration for 16S amplicon sequencing, can you keep using it, or do you have to rebuild it every time your lab does a run? By expanding the usual single-condition calibration into a multi–run, multi–instrument, multi–primer design and then fitting a hierarchical model, we were able to separate variability that comes from the taxon–primer pair itself from variability that is introduced by day-to-day practice and by reagents. The key messages from the Results are: (i) the core log-ratio linear pattern is very stable; (ii) ordinary run-to-run and even instrument changes only add a small amount of spread; (iii) primer-lot changes add the largest, structured shifts; and (iv) this is predictable enough that a simple QC rule can protect laboratories from reusing non-portable calibrations.

Our findings support a *conditional* view of portability. PCR-bias calibration is portable across runs and across thermocyclers *when* the underlying amplification chemistry as experienced by the primer–template pair is the same. This is why we saw high reproducibility of slopes within runs and only modest run-level variance components: the same master DNA pool, same primer lot, and same basic protocol produce almost the same slopes even on different days [13, 14]. Likewise, the two

thermocyclers we tested did not induce large shifts, suggesting that, for modern instruments with comparable ramp rates and accurate hold temperatures, the "instrument effect" is real but small enough that it can be absorbed into the uncertainty of a reused calibration [15].

By contrast, once we changed primer lots, slopes for a non-trivial fraction of taxa moved enough to break our portability threshold [16]. This is consistent with how 16S amplification works in practice: even small lot-to-lot differences in oligo concentration or purity can change effective primer availability, and those changes propagate most strongly to taxa already sitting near the edges of primer complementarity. In other words, portability fails exactly where microbiome researchers care about it most — in the tails of the community. This is why we do not recommend assuming a "calibrate once, reuse forever" model.

It is worth emphasizing why primer lots, rather than thermocyclers, emerged as the dominant technical factor. Thermocyclers mostly affect *how* fast the reaction reaches the intended temperature profile. If both instruments reach 55°C for annealing and 72°C for extension with reasonable accuracy, then the environment seen by the primer–template complex is nearly identical, and taxon-specific slopes stay similar. Primer lots, on the other hand, affect the *molecular participants* directly: slight changes in the amount of forward vs. reverse primer, low-level degradation, or differences in purification can change the kinetics of binding for some templates more than others. Because our model quantifies bias as a slope over multiple cycles, and because small per-cycle differences accumulate over 30–35 cycles, primer-lot shifts are naturally amplified and become visible [17].

A practical corollary is that primer metadata should be treated as first-class experimental metadata, on par with sequencing run IDs [18]. Many public 16S datasets list the primer sequence but not the lot, making it difficult to decide whether a single published calibration can be reused. Our results suggest that groups who want to share calibrations should also share primer-lot identifiers, or else encourage recipient labs to run the one-sample QC check we proposed.

One of the most useful outcomes of this work is that we can now propose a lightweight QC gate that is aligned with the calibration model itself. Instead of asking labs to eyeball amplification curves or re-estimate the full hierarchical model for every run [19], we ask them to (i) include the pooled community at one or two cycle numbers on every run, (ii) predict those compositions from the archived calibration, and (iii) compute an Aitchison distance. If the distance is below 0.15 (or a stricter 0.10 for rare-taxa studies), the run is deemed portable and the calibration can be reused. If not, the run is flagged and a new multi-cycle calibration is warranted. This is simple enough to automate in an R or Python script and adds only a single extra library to each run.

Importantly, this QC gate is *compositional*, i.e. it uses the same log-ratio framework as the calibration. Many sequencing QC steps still rely on count-space or percentage-space comparisons, which can hide the kinds of small, taxon-specific shifts we saw here. By keeping QC and calibration in the same space, we preserve sensitivity to exactly the effects that matter.

Our portability results also have implications beyond the single lab. A growing number of microbiome studies are reanalyzed years later, in meta-analyses or in method benchmarking, and calibration parameters could, in principle, travel with those datasets. Our results suggest a cautious approach: calibration parameters learned from a study are likely reusable by another study *only* if (i) the primer lot is the same or demonstrably equivalent, and (ii) the receiving study is willing to run a small verification panel. Otherwise, the safest route is to treat the published calibration as a prior or starting point, and update it with local multi-cycle data. This is where the hierarchical model we used is helpful, because it naturally accommodates adding new "groups" (new runs, new

instruments, new primer lots) without discarding what was learned before.

5. Conclusion

This study showed that PCR-bias calibration for 16S amplicon sequencing can be reused, but only under the right conditions. When the same pooled DNA, the same protocol, and especially the same primer lot are used, the taxon-specific "slopes" that describe how relative abundances change with PCR cycles are very reproducible, even across different days and even on a different thermocycler. That means the basic four-step idea — make a pooled community, amplify it at several cycle numbers, fit a log-ratio linear model, and then use those parameters to correct real samples — actually holds up in realistic lab settings.

What breaks portability is changing primer lots. Lot-to-lot differences in the very same 16S primer pair were the biggest, most systematic source of slope changes, and because bias accumulates over 30–35 cycles, those changes are big enough to make an old calibration unsafe to reuse. The good news is that this is easy to monitor: if the lab puts one pooled-community sample on each run and checks how close its composition is to what the old calibration predicts (using Aitchison distance), the lab can decide automatically whether to reuse or to recalibrate. So the final message is: calibrate when the primer lot changes; otherwise reuse, but always run the small compositional QC.

References

[1] Golob, J. L., Margolis, E., Hoffman, N. G., & Fredricks, D. N. (2017). Evaluating the accuracy of amplicon-based microbiome computational pipelines on simulated human gut microbial communities. *BMC Bioinformatics, 18*(1), 283.

[2] Poretsky, R., Rodriguez-R, L. M., Luo, C., Tsementzi, D., & Konstantinidis, K. T. (2014). Strengths and limitations of 16S rRNA gene amplicon sequencing in revealing temporal microbial community dynamics. *PloS one, 9*(4), e93827.

[3] McLaren, M. R., Nearing, J. T., Willis, A. D., Lloyd, K. G., & Callahan, B. J. (2022). Implications of taxonomic bias for microbial differential-abundance analysis. bioRxiv, 2022-08.

[4] Silverman, J. D., Bloom, R. J., Jiang, S., Durand, H. K., Dallow, E., Mukherjee, S., & David, L. A. (2021). Measuring and mitigating PCR bias in microbiota datasets. *PLoS Computational Biology, 17*(7), e1009113.

[5] Harrison, J. G., John Calder, W., Shuman, B., & Alex Buerkle, C. (2021). The quest for absolute abundance: The use of internal standards for DNA-based community ecology. *Molecular Ecology Resources, 21*(1), 30-43.

[6] Brooks, J. P., Edwards, D. J., Harwich Jr, M. D., Rivera, M. C., Fettweis, J. M., Serrano, M. G., ... & Buck, G. A. (2015). The truth about metagenomics: quantifying and counteracting bias in 16S rRNA studies. *BMC Microbiology, 15*(1), 66.

[7] Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017, July). On calibration of modern neural networks. In *International Conference on Machine Learning* (pp. 1321-1330). PMLR.

[8] Mannheimer, S., Sterman, L., & Borda, S. (2016). Discovery and reuse of open datasets: An exploratory study. *Journal of eScience Librarianship, 5*(1), e1091.

[9] van Pelt-Verkuil, E., Van Belkum, A., & Hays, J. P. (2008). *Principles and Technical Aspects of PCR Amplification*. Dordrecht: Springer Netherlands.

[10] Minderer, M., Djolonga, J., Romijnders, R., Hubis, F., Zhai, X., Houlsby, N., ... & Lucic, M. (2021). Revisiting the calibration of modern neural networks. *Advances in Neural Information Processing Systems, 34*, 15682-15694.

[11] Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017, July). On calibration of modern neural networks. In *International Conference on Machine Learning* (pp. 1321-1330). PMLR.

[12] Despagne, F., & Massart, D. L. (1998). Neural networks in multivariate calibration. *Analyst, 123*(11), 157R-178R.

[13] Lorenz, T. C. (2012). Polymerase chain reaction: basic protocol plus troubleshooting and optimization strategies. *Journal of Visualized Experiments: JoVE*, (63), 3998.

[14] Asif, S., Khan, M., Arshad, M. W., & Shabbir, M. I. (2021). PCR optimization for beginners: a step by step guide. *Research in Molecular Medicine, 9*(2), 81-102.

[15] Gupta, S. V. (2012). *Measurement Uncertainties: Physical Parameters and Calibration of Instruments*. Springer Science & Business Media.

[16] Weninger, M., Gander, E., & Mossenbock, H. (2021). *User Guidance and User Behavior*. Sworn Declaration, 135.

[17] Lee, A., & Wong, E. (2009). Optimization and the robustness of BOX A1R PCR for DNA fingerprinting using trout lake E. coli isolates. *Journal of Experimental Microbiology and Immunology, 13*, 104-11.

[18] Task Group on Data Citation Standards and Practices, C. I. (2013). Out of cite, out of mind: The current state of practice, policy, and technology for the citation of data. *Data Science Journal, 12*(0), CIDCR1-CIDCR75.

[19] Bosch, J., Armstrong, R., Bickerton, S., Furusawa, H., Ikeda, H., Koike, M., ... & Yamanoi, H. (2018). The hyper suprime-cam software pipeline. *Publications of the Astronomical Society of Japan, 70*(SP1), S5.

How to cite this article: Sayan Mukherjee (2023). Portability of PCR-Bias Calibration in 16S Amplicon Sequencing Across Runs, Instruments and Primer Lots. *Bulletin of Computer and Data Sciences*, 4(2), 18-29. DOI: 10.71448/bcds2342-2

Received: 09/05/2023 **Revised:** 22/06/2023 **Accepted:** 01/07/2023 **Publish:** 30/08/2023