Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Family-specific structure in infant gut viromes enables closed-set mother-child matching

Anmin Wang
the Chinese Academy of Sciences, University of Science and Technology of China
Lauren Lai
David Geffen School of Medicine, University of California, Los Angeles (UCLA)
Colin Patrick Johnson
David Geffen School of Medicine, University of California, Los Angeles (UCLA)
Xiaoyue Yang
The First Affiliated Hospital of University of Science and Technology of China (USTC), University of Science and Technology of China
Chao Ye
yechao850606@ustc.edu.cn

The First Affiliated Hospital of University of Science and Technology of China (USTC), University of Science and Technology of China

Article

Keywords: infant gut virome, bacteriophage, mother-infant sharing, family structure, consensus-sequence concordance, twins, closed-set matching

Posted Date: November 3rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10978238/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

The early-life gut virome differs so strongly between individuals that the variation is usually treated as noise. We show it is structured by family, specifically enough that a child's virus set points to their own mother. Across 2,664 catalogue metagenomes, 142 twin libraries and 284 shotgun metagenomes re-quantified from raw reads, we ranked 32 candidate mothers by viral overlap with each of 125 child samples from 32 families. Using all detected vOTUs, the true mother ranked first for 67 of 125 samples (53.6%; median rank 1), against a median sample-level accuracy of 2.4% after permuting maternal labels across families. The mean within-family accuracy was 52.0% (family bootstrap 95% confidence interval 37.2% to 66.6%), and the true mother ranked first at least once in 22 of 32 families. Specificity comes from where the background is empty, not where the signal is large: the own-mother excess is largest for viruses of intermediate prevalence (+ 0.64) and small for the rarest (+ 0.09), yet rare viruses discriminate best because unrelated pairs share almost none. The same asymmetry appears independently in the twin cohort, where rare-virus sharing separates co-twins from unrelated pairs at AUC 0.997. Matching is not merely taxonomic: median agreement across cohort-variable consensus positions in shared vOTUs was 1.000 for mother-child pairs against 0.768 for unrelated pairs, positive in all 24 evaluable families. These findings support family structure in infant viromes, with additional within-vOTU consensus-sequence concordance in the evaluable subset of shared viruses. The comparisons are closed-set and do not resolve transmission direction or separate maternal acquisition from continued shared exposure.

Introduction

The gut virome of an infant is assembled within months of birth, and where it comes from remains contested [1, 2, 3]. By the time these communities can be measured they already differ markedly between individuals, a pattern seen across early-life catalogues and cohort meta-analyses [4, 5]. The prevailing description is progressive environmental acquisition: viruses arrive from the surroundings, from diet, and from close contacts, and the community gradually comes to resemble that of adults [1, 3]. Read as a sampling model, that description makes a testable prediction. Under a simple population-pool sampling model the draw is weighted toward what is abundant in that pool, so the viruses an infant carries should be enriched for those common among surrounding adults. Environmental acquisition need not take this form — household, diet and geography can make a population-rare virus locally common — but the population-pool version is the one that makes a sharp prediction.

A second process, not mutually exclusive with the first, is disproportionate exposure to a small number of close contacts, as reported for bacteria [6] and, more recently, for viruses [7]. This additionally predicts family-specific sharing, including viruses that are uncommon in a wider reference population or among candidate family members. The two processes can operate together; they differ in what they predict about prevalence and about family structure, not in being alternatives.

These predictions differ in a quantity that public data can address — the population prevalence of the viruses an infant carries. We evaluated it in the Unified Human Gut Virome catalogue [8], then asked

whether whatever structure we found was family-specific, whether it pointed at the mother, and whether it persisted.

Two features of the data required care. Detection of a virus depends on how deeply a sample was sequenced and on how many viruses it carries in total, and infants differ from adults on both. Any statistic summarizing "the prevalence of what this sample carries" is therefore susceptible to differences that have nothing to do with age. We show below that the natural version of this statistic is substantially biased under a null in which no age effect exists, and that the direction of the bias depends on analysis choices that appear innocuous. Every comparison we report is consequently referenced to its own matched null rather than to a nominal value.

Results

A child's virome ranks their own mother first among thirty-two candidates

For each of 125 child samples we computed the Jaccard overlap between the child's full detected vOTU set and each of the 32 mothers' detected sets pooled across her timepoints, without filtering either set by maternal prevalence. The child's own mother scored strictly highest for 67 samples (53.6%) and had a median rank of 1 (Fig. 1a,b). Five samples tied with the best unrelated mother; all five shared no detected vOTUs with any mother. The primary analysis assigns every tied candidate the worst rank in its tied group, making top-one accuracy conservative with respect to ties. Because samples repeat within families, sample-level accuracy is descriptive: averaging each family's sample-level accuracy gives 52.0%, with a family bootstrap 95% confidence interval of 37.2% to 66.6%. Drawing one sample per family gives a median of 53.1% and a central 95% draw interval of 43.8% to 62.5% across 2,000 draws. The true mother ranked first for at least one sample in 22 of 32 families and for every sample in 12. Poorer ranks were associated with fewer detected vOTUs in the child (Fig. S1d).

The null permutes the mother-to-family assignment as a single bijection across families, preserving all repeated child samples, maternal sets and observed score ties. Top-one accuracy was 53.6% against a null median of 2.4% (central 95% permutation interval 0.0% to 8.8%); top-three accuracy was 66.4% against 8.0%, and top-five accuracy was 72.8% against 13.6% ($P = 1/2{,}001$ for each at $B = 2{,}000$; Fig. 1c, Fig. S1a,b,e,f). Under a uniform random assignment, a unique top-ranked mother has probability 1/32 of receiving the true label; maternal richness does not change that expectation. With the conservative tie rule, the exact mean random-label top-one accuracy across these samples is 3.0%, whereas 2.4% is the empirical null median. The median of the permutation-specific median ranks was 21, because ties are assigned their worst rank. Twenty-three samples placed the true mother last, and rank was associated with child detected richness (Spearman $\rho = -0.49$, Fig. S1d).

Rare viruses supply specificity, common viruses supply effect size

Splitting each child's detected set by how many of the 32 mothers carry each virus separates the magnitude of the excess from its pairwise specificity (Fig. 1d). The excess sharing with one's own mother is present in all three prevalence strata and significant in each against its own family-label permutation ($P \leq 0.0005$, the floor at $B = 2{,}000$; Fig. S2a-d), but it is not largest among rare viruses. It is + 0.639 for uncommon viruses, + 0.412 for common ones, and only + 0.089 for rare ones.

What is particular to rare viruses is the baseline. An unrelated mother shares none of a child's rare viruses in 73% of pairs, against 34% for uncommon and 2% for common viruses; median sharing with an unrelated mother is 0.000, 0.111 and 0.520 across the three strata. A rare virus in common between a child and a woman is therefore far more specific to a related pair than a common one, even though the own-mother excess is larger for common viruses. Rare viruses contribute the specificity of a family match because unrelated-pair sharing is frequently zero, not because the signal is strong. This prevalence decomposition is distinct from the primary ranking, which uses all detected vOTUs. Figure S1c compares tie rules applied to the same full-set scores and is not a comparison of prevalence cutoffs.

The supplementary prevalence-restricted analysis addresses a different scoring definition from the primary full-set analysis. It restricts only the child's vOTU set to viruses carried by at most 30% of candidate mothers, while leaving each maternal set unfiltered. Its archived top-one accuracy is 58.4%, compared with 53.6% for the primary full-set score. Within the archived restricted analysis, excluding the child's own mother from the prevalence reference leaves top-one accuracy at 58.4%, and excluding both the assigned and candidate mothers gives 56.0% (Table S1). These values must not be treated as alternative estimates from the same scoring pipeline. The leave-out comparisons require filters that follow the permuted maternal assignment when constructing their nulls; their archived P values are retained for provenance and do not support a claim of independently verified robustness. All identification results remain closed-set comparisons among these candidate mothers.

The twin cohort shows the same asymmetry from the other direction. Co-twins share three times more of their uncommon viruses than their rare ones in absolute terms (medians 0.470 versus 0.144), yet rare viruses separate twin pairs from unrelated pairs almost perfectly, AUC 0.997 against 0.943 for uncommon viruses (Fig. 1e). Magnitude and discriminability move in opposite directions across the prevalence range, and only the second is what an identification claim rests on.

Shared viruses show greater consensus-sequence concordance within families

Detection at the level of a viral operational taxonomic unit cannot distinguish a virus a child acquired from their mother from an unrelated virus of the same kind. We therefore re-quantified the cohort from raw reads against the 309 viral genomes that reach tenfold coverage in both a child and a mother (Fig. 2e), called a consensus base at every position covered at least fivefold with at least 80% agreement, and compared child and mother base by base (Fig. 3).

Most positions are invariant across the whole cohort, so agreement everywhere is uninformative: across all jointly called positions an unrelated mother already matches at 0.991, against 1.000 for the child's own mother (Fig. 3f). Restricting to positions where at least two bases occur among all individuals carrying that genome — a median of 190 of 9436 called positions per genome (1.9% per genome, 2.6% of all called positions pooled), defined before any own-versus-unrelated split — separates the arms cleanly. Agreement with the own mother is 1.000 (IQR 0.919-1.000) against 0.768 (IQR 0.606–0.920) with unrelated mothers across 5,011 own and 88,181 unrelated comparisons on 241 genomes (Fig. 3a, Fig. S3a,c,d,f). 52.4% of own-mother comparisons are identical at every shared variable position, against 8.0% of unrelated comparisons (Fig. 3d).

The contrast is a property of families rather than of a few deep genomes. Of the 79 child samples with sufficient coverage, 78 showed higher median agreement with the matched mother than with unrelated mothers (Fig. 3b). Because samples repeat within children, inference was performed after aggregating to one value per family: the own-minus-unrelated difference is + 0.216 and positive in all 24 families, against a null of -0.009 [-0.035, + 0.016] when maternal labels are permuted across families ($P \leq 0.0005$, the resolution floor at B = 2,000; Fig. 3c). The effect is spread across genomes rather than carried by a few, and restricting the positions raises the same per-child quantity rather than changing its sign (Fig. S3b,e). To assess sensitivity to coverage, the comparison was repeated within quartiles of joint depth. Greater own-mother concordance was observed in every quartile (Fig. 3e); this descriptive stratification does not exclude all depth-related bias.

This analysis provides within-vOTU consensus-sequence concordance beyond co-detection: for viruses covered deeply enough to call a consensus, matched pairs agree more closely at cohort-variable positions than unrelated pairs do. It does not establish direction: a maternal sample taken earlier predicts the child no better than a later one (11 of 32 families, Fig. 2c), so these data are equally consistent with transfer to the child and with continued shared exposure within the household.

Infants carry viruses that are rare among adults

For each sample we computed the median prevalence, among adults of the same study, of the viral operational taxonomic units (vOTUs) that sample carried, scoring adults leave-one-out, and compared infants against adults by rank. Values below one half would indicate that infants carry viruses rarer among adults, but the uncalibrated statistic has no stable null expectation. Drawing pseudo-infants at random from adults only, where the true answer is one half by construction, returned 0.380 and 0.398 in the two infant-containing studies under one inclusion rule, and 0.937 and 0.685 under another that differed only in whether vOTUs carried by a single adult were retained. The bias is structural: focal samples are scored against the full reference panel while reference samples are scored leave-one-out, and a median over a sparse carried set is unstable. Raw values from this statistic cannot be compared to one half, and the two inclusion rules disagree about the direction of any such comparison.

We therefore interpret each variant only relative to its own matched null (Fig. 4a,b,e), and the whole grid of variants is shown against those nulls (Fig. 4c). Against pseudo-infants matched to real infants on log detection richness and log sequencing depth — matched medians 27 versus 26 and 127 versus 138 vOTUs, with depth equal to two decimal places — infants fell below the matched null in both studies and under both inclusion rules, by − 0.161 and − 0.120 (Tett 2019 [9], study 2) and − 0.035 and − 0.079 (Carter 2023 [10], study 21) for the two rules, with empirical P from 1.0×10^{-4} to 0.0038 across the four comparisons at 10,000 permutations, three of them at the resolution floor. The conclusion does not depend on which inclusion rule is chosen.

Across 24 cells crossing three inclusion rules with four matching schemes, every observed value lies below its own null (Table S2, Fig. S4a), and the unmatched calibration cells show the bias changing sign with the inclusion rule (Fig. S4b). The grid is a sensitivity analysis of sign and calibration rather than a family of primary hypothesis tests. In 21 of the 24 cells the null is a proper distribution and the largest empirical P among them is 0.013; the remaining 3, all using richness-only matching in the Carter 2023 study, returned a single distinct null value because the nearest-neighbour match there is unique, so their P values are not interpretable (Table S2). Unmatched draws, where the correct value is one half, carry a bias of − 0.12 to + 0.32 whose sign depends on the inclusion rule, which is why matched nulls are used throughout.

The same statistic applied to a wider age contrast, using per-sample richness- and depth-matched nulls across the 943 samples in the seven studies with recorded ages, did not separate children under two from adults (permutation P = 0.084, B = 300); with 300 draws this analysis is underpowered and is reported only to indicate that the effect above is specific to the two infant-containing studies rather than a general age gradient.

Rare viruses are shared within families and not between them

If the rare viruses infants carry come from specific individuals, they should be shared by people in close contact. The catalogue's virus-like-particle layer includes a Malawian twin cohort, providing 142 samples from 40 individuals in 20 families with co-twin pairing. Runs from the same individual were collapsed to one virome by union before any pairwise comparison, so the units here are individuals, not runs. This layer uses a different enrichment and detection regime from the bulk catalogue, and is used as orthogonal validation of family structure rather than as a comparable effect-size estimate.

Stratifying by population prevalence and comparing Jaccard similarity between individuals (Fig. 1e, Fig. 2b), co-twins shared 14.4% of their rare viruses (≤ 10% prevalence) against a median of exactly zero for unrelated children — a difference of + 0.144 against a null of + 0.003 [− 0.003, 0.012] under permutation of family labels across individuals, P ≤ 0.0005, the floor at B = 2,000. For uncommon viruses (10–30%) the corresponding values were 0.470 and 0.061, difference + 0.410, P ≤ 0.0005 (B = 2,000). Too few co-twin pairs carried enough common viruses to test that stratum. The rare stratum is the informative one, though the separation is one of degree: 48.4% of unrelated pairs share at least one rare

virus, but their sharing never exceeds a Jaccard of 0.058, whereas every co-twin pair shares some and 19 of 20 exceed the entire unrelated range. Rare-virus sharing is therefore strongly enriched within families.

Children share more viruses with their own mother than with unrelated mothers

Family structure does not identify which family member matters. We therefore quantified an independent mother-infant cohort from raw reads — 44 families with maternal samples across gestation, birth and three months, and child samples across birth, 14 days, and one, two and three months, with no biosample overlap with the catalogue. Of 286 runs, 284 produced output; 38 fell below three million reads and were excluded, of which 29 were birth meconium samples (Fig. 2d), so the birth timepoint is under-represented by biology rather than by choice. The analysable set is then 43 families with a usable maternal virome and 33 with a usable child virome, intersecting at 32 families contributing 131 child samples, of which 125 pass a requirement of at least five detected viruses; delivery mode is recorded for 41 of the 44 enrolled families and for all 32 analysable ones.

For each child sample we measured the fraction of that child's detected vOTUs also detected in a given mother, and compared the value for the child's own mother against the mean over all unrelated mothers. Children shared a larger fraction with their own mother (Fig. 1a,d): the contrast was positive in 25 of 32 children, and in 88 of the 125 child samples those children contributed (median 4 samples per child, range 2 to 5). Because observations repeat within children and mothers, inference was performed at the child or family level: the 125 own pairs and 3,875 unrelated pairs reuse the same 32 children and 32 candidate mothers, and each child contributes two to five samples, so per-sample values are reported for description only. Every reported statistic uses the 32 mothers with a paired child; the 43 families with a usable maternal virome are the enrolment set, not an analysis set (Table S3). The primary estimand is the child-level excess, + 0.156; permuting the mother-to-family assignment gave a null of − 0.002 [− 0.017, 0.002] and $P \leq 5.0 \times 10^{-5}$, the resolution floor at 20,000 permutations. The corresponding value pooled over samples, + 0.110, is reported for description only. Restricting each mother to a single deepest run rather than pooling her timepoints preserved the effect. The excess is not confined to one age: splitting the pairs by the child's age gives own-minus-unrelated differences of + 0.33 at birth, + 0.29 at 14 days, + 0.16 at one month, + 0.24 at two months and + 0.24 at three months (Fig. 2a), the birth value resting on the 7 children whose meconium samples passed the depth threshold. Rarefying every mother to an equal detected-set size of 194 vOTUs, computed at the child level so that it refers to the same estimand as the headline, reduced the excess from + 0.156 to + 0.069 [0.049, 0.086]. Rather more than half of the raw effect therefore reflects how much of each mother was observed, and the remainder does not. Regressing the sharing fraction on own-mother status with maternal detected-set size and maternal depth as covariates left the own-mother coefficient at + 0.161 (standard error 0.011); this regression treats samples as exchangeable and is reported as a covariate check rather than as a significance test.

vOTUs shared with the own mother were rarer in the population than those shared with unrelated mothers (0.279 versus 0.465). This comparison is mechanically biased, since coincidental sharing with an unrelated mother requires a widespread vOTU. Permutation of mother labels absorbs exactly that construction: the permuted value rises to 0.465, matching the unrelated arm, while the observed 0.279 lies far below the permutation null [0.419, 0.488], $P \leq 0.0005$ ($B = 2{,}000$), with 22 of 27 children individually below their own unrelated mean, across the 95 child samples in which at least five rare viruses were detected — the subset in which a rare-fraction prevalence is defined, drawn from the 32 children and 125 samples of the full analysis.

No association was apparent with the coarse batch variables available in the public metadata: all runs share one instrument model and one submission date, but extraction, plate, lane and index effects cannot be evaluated from these records. Permuting mother labels only within library-processing block, which preserves processing order while destroying kinship, gave $P \leq 0.0005$, the floor at $B = 2{,}000$. As a negative control, the excess did not differ by child sex (+ 0.122 male versus + 0.108 female, $P = 0.85$, $B = 5{,}000$).

Exploratory delivery-mode differences in own-mother excess

Delivery mode offers a partial discriminator between explanations that family structure alone cannot separate. Vaginal delivery exposes the infant to maternal microbiota directly; caesarean delivery largely bypasses that exposure while leaving household sharing intact. It is only partial, because caesarean birth also alters early colonization, antibiotic and hospital exposure, feeding, and carries the maternal indication for surgery.

In this exploratory subgroup analysis the family-level own-mother excess was larger among vaginally delivered families than among the four caesarean-delivered families: median + 0.182 across 28 vaginal families against + 0.001 across the 4 caesarean families (Fig. S5a), an exact family-level permutation over all 35,960 assignments giving one-sided $P = 0.015$. With four caesarean families this contrast is treated as suggestive. Dropping each caesarean family in turn left the contrast between + 0.180 and + 0.182.

Caesarean children in this cohort carried far fewer detected viruses (median 37 versus 88 at equal sequencing depth), and their sharing with unrelated mothers was correspondingly depressed, so the raw contrast is confounded with detection. Three sensitivity analyses evaluate this detection imbalance; the sharing ratios and richness-matched subset are illustrated in Fig. S5b. Expressed as a ratio of own to unrelated sharing, which is insensitive to a uniform depression of both terms, vaginal families gave 2.14 against 0.53 (exact permutation over all family assignments, $P = 0.015$). Restricting to a vaginal subset matched to caesarean children on detection richness (median 40 vOTUs in both arms after matching, against 37 in the unmatched caesarean group) gave 2.18 against 1.00 (exact permutation, $P = 0.028$). Rarefying every child to ten detected viruses gave a vaginal excess of + 0.153 against + 0.000 for caesarean children; the quoted ranges [0.113, 0.204] and [− 0.002, 0.001] are technical intervals over

resampling draws, conditional on these families, and are not confidence intervals over families. The contrast is therefore not explained by these three detection summaries alone, though detection differences combined with delivery-associated biology remain possible.

Exploratory longitudinal and temporal analyses

An exploratory longitudinal analysis did not yield a stable retention claim. Among the 32 families with at least two post-birth child timepoints, viruses also present in the mother were retained more often than others within the rare stratum (family-level means 82.5% against 45.2%, higher in 15 of the 19 families with both arms represented) but not within the uncommon or common strata. The aggregate test is unstable: with a prevalence caliper enforced, which retains only 315 of 888 maternal viruses, the paired test across children gave Wilcoxon $P = 0.11$ while family-level permutation gave $P = 0.026$ ($B = 2{,}000$) with the observed value at the null's upper bound. A formal maternal-status by prevalence-stratum interaction moved from $P = 0.66$ to $P \leq 0.0005$ ($B = 2{,}000$ in both) depending only on whether gestational maternal samples entered the definition of a maternal virus. We therefore make no claim about prevalence-specific retention. The underlying values, with standard errors computed across families rather than across vOTUs, are shown in Fig. 4d.

Maternal samples taken earlier (Fig. 2c) did not predict a child's first virome better than samples taken later (early 0.243 versus late 0.176, but higher in only 11 of 32 families, Wilcoxon signed-rank $P = 0.57$). Direction of transfer is therefore not established, and continuing shared exposure fits these observations as well as acquisition at birth does. The delivery-mode contrast is insufficient to distinguish birth-route exposure from household and clinical factors: caesarean families share a household too, and the contrast rests on four families confounded with the clinical correlates of caesarean birth; it does not demonstrate transfer at delivery.

Discussion

Across three data layers, how common a virus is in the population determines what its sharing can be used for. The practical consequence is that a child's virome identifies their own mother out of thirty-two candidates in over half of samples, + 51 percentage points above a permuted-label baseline, using all detected vOTUs in the primary Jaccard score. And for viruses covered deeply enough to call a consensus, the agreement extends below the vOTU level: matched pairs agree at cohort-variable positions where unrelated pairs differ. This is within-vOTU consensus-sequence concordance; it does not establish complete strain identity or transmission. In both catalogue studies containing infants, infants carry viruses rarer among adults than matched expectation; rare viruses are strongly enriched for sharing within families; and children share an excess of viruses with their own mother that persists after adjustment for maternal detection opportunity and detection richness, and when maternal labels are permuted within library-processing block. The three lines come from a bulk catalogue layer, a virus-like-

particle layer and reads we quantified ourselves, which differ in detection regime and are therefore triangulation rather than a single graded measurement.

A population-pool sampling model does not readily explain this. Drawing in proportion to population abundance predicts enrichment for common viruses, and we observe the opposite in both studies where infants are present. The pattern is compatible with early exposure structured by a small number of close contacts, whose personal viruses are necessarily uncommon at population scale, but a household-local pool would produce much the same signature.

Our result sits alongside a small existing literature on mother-infant virus transfer. Strain-resolved work on bacteria established that maternal transmission dominates early colonization in the same cohort we re-quantified [6], and a longitudinal study of mother-infant pairs reported virus transmission during pregnancy and early life using assembly and strain-level methods [7]. Our contribution is not the existence of mother-child sharing but its prevalence dependence: the excess is present in every cohort-mother prevalence stratum and largest for uncommon viruses, while rare viruses supply the pairwise specificity, and it is recovered at vOTU resolution without assembly, in a catalogue-scale reference frame that makes population prevalence measurable in the first place [8]. That work analysed virus-like-particle and total metagenomes in 30 mothers and 32 infants and reported both maternal sharing and post-birth cohabitation effects [7]; where it asked whether transfer occurs, we ask which part of the virome it acts on. The original report on the twin cohort we use here found that co-twins resembled each other most, and that twins resembled unrelated individuals somewhat more than they resembled their own mothers [11] — the opposite ordering to our mother-child result. These studies measure different components of early-life virome sharing. That comparison was whole-virome similarity in a Malawian virus-like-particle survey, ours is a prevalence-stratified sharing fraction in a different cohort quantified from bulk reads, and our own data reproduce the first half of their result: co-twin sharing is the largest effect we measure. A twin study of 28 infant pairs and their mothers likewise found co-twins most similar and reported no effect of delivery route on mother-infant sharedness [12], consistent with our four-family caesarean contrast being too small to resolve. Differences in prevalence weighting may contribute: whole-virome similarity is dominated by common viruses, which unrelated people share almost as readily as relatives do, so they carry a large excess with little specificity; rare viruses carry a smaller excess against an almost empty background, and only the second can identify anyone. That is a testable claim rather than a settled one: it predicts that recomputing their maternal comparison on the rare fraction alone would reverse its ordering, which the four quantified mothers in that cohort are too few to check. Reports that the infant virome assembles stepwise and is modulated by feeding [1], that much of its diversity is novel relative to adult references [2], and that early virome and bacteriome dynamics are coupled [3] are all consistent with a rare, individual-specific early fraction, but none of them measures population prevalence directly.

Nine limits bound the interpretation. First, the mother is confounded with the household: these data cannot separate maternal transfer from shared environment, and doing so requires sampling fathers, siblings or the household itself. The delivery-mode contrast is the closest available discriminator and

rests on only four caesarean families. The data support family-structured sharing in which mother-child pairs exceed unrelated pairs; they do not identify the mother rather than the household as the source. Second, the longitudinal analysis is exploratory only: its aggregate form is unstable and its stratified form is sensitive to an unrelated analytic choice, so no retention claim is made. Third, the exclusion of shallow meconium samples removes the earliest acquisition window, so the interval in which colonization actually occurs is the interval these data cover least well. Fourth, the twin cohort's clinical background is not available. It was collected to study severe acute malnutrition, roughly half its pairs were discordant for it and treated, and neither that status nor zygosity appears in the public metadata, so the family-sharing result cannot be stratified by either. Both effects would have to act within families to generate the contrast we report, but this cannot be verified from these data. Fifth, only one of the three data sources contains mothers. The twin cohort deposits maternal samples for 9 of 20 families, but only 4 of these were quantified in the catalogue's abundance tables, too few to serve as a second test of the mother-child result; that result therefore rests on a single cohort, and the twin data speak only to family structure. Sixth, the three layers are not measured alike: the mother-infant cohort is bulk shotgun sequencing in which viruses are called at a coverage threshold and detection is strongly depth-dependent, whereas the twin layer is a virus-like-particle preparation. Effect sizes are therefore not comparable across layers, and only the direction of each contrast is. Seventh, the prevalence strata are defined within this cohort rather than at catalogue scale: only 490 of the 2,843 viruses these children carry appear in the catalogue-adult prevalence table at all, so a catalogue-scale split cannot be computed for them and rare here means rare among these 32 mothers. Eighth, identification is demonstrated within a closed set of 32 known mothers; nothing here shows the ranking would survive against an open population, which is the comparison any applied use would require. Ninth, the consensus-concordance comparison covers only the 309 of 2,843 viruses that reach tenfold coverage in both members of a pair, a minority selected for abundance, and supports greater within-vOTU consensus-sequence concordance among evaluable shared viruses. It does not establish whole-genome strain identity, reconstruct haplotypes, or extend sequence evidence to every shared vOTU.

One methodological note has wider application. Where a summary statistic depends on detection depth and set size — as most prevalence, sharing and overlap statistics do — comparing it against a nominal value is not sufficient, because the statistic used here was biased in either direction depending on an inclusion rule that appears inconsequential (Table S2). A negative-control resampling design bounds label-driven artefacts, though adult pseudo-samples need not reproduce infant detection biology in every respect.

Methods

Data sources

Catalogue analyses used the Unified Human Gut Virome resource [8]: 168,536 vOTUs with host-genus assignments and read-mapping relative abundances across 2,736 bulk metagenomes from 21 studies, plus a virus-like-particle layer. Published detection in that resource applies asymmetric coverage-breadth

thresholds, 75% for viral contigs and 50% for prokaryote genomes, with sequential mapping. Two of the 21 studies contain samples from children under three years, and together account for every such sample in the resource. Age was taken from the read-mapping sample metadata where present and from the source biosample metadata otherwise, giving 104 samples under three years in these two studies; 103 carried at least five detected vOTUs, and 99 carried at least five vOTUs also present in the adult reference panel and entered the prevalence comparison (32 in Tett 2019 and 67 in Carter 2023, against 81 and 180 adult samples). Reporting only one of the two age fields undercounts this set.

Twin analyses used the Malawian cohort within the virus-like-particle layer, keyed as $PRJEB9818_ < run >$ [11]. That cohort was assembled to study severe acute malnutrition, and its published design comprises monozygotic and dizygotic pairs of which roughly half were discordant for malnutrition and received therapeutic food [11]. Neither zygosity nor nutritional status is recoverable from the public metadata: the ENA sample attributes for these runs carry only environment, geography, collection date and sequencing platform, and the $disease$, $host_phe \neg ype$ and $host_status$ fields are empty for all 503 runs. The twin analyses are therefore unstratified by either variable, and are used only for the within-family versus between-family contrast, in which both twins of a pair share whatever nutritional exposure the family had. Multiple runs from one individual were collapsed to a single detected set by union before pairwise comparison, so pairs are between individuals; permutation shuffles family labels across individuals. Family identity is encoded in the source biosample metadata as $Ma_F < family > T < tw \in >_< date >$; run-to-sample links were obtained from the European Nucleotide Archive, yielding 142 runs from 40 individuals in 20 families.

The mother-infant cohort (PRJNA475246) [6] was quantified from raw reads. Gzip-compressed FASTQ files were downloaded from the European Nucleotide Archive by FTP; each file was verified for archive integrity and for four-line record structure, and truncated files were treated as fetch failures and retried. Reads were used as deposited: no adapter or quality trimming, no complexity filtering, no duplicate removal and no host-read depletion were applied, because the reference is a closed catalogue of viral and prokaryote genomes, so human-derived and low-quality reads fail to align rather than being miscounted, and because this keeps the quantification comparable to the published catalogue pipeline that we mirror. Libraries with fewer than three million reads were excluded before mapping. Paired reads were mapped with bowtie2 v2.5.5 [13] at default sensitivity and default multi-mapping behaviour, in which each read is reported at a single best-scoring location chosen pseudo-randomly among equal-scoring ties, against the vOTU representative sequences ($\bowtie 2 - p4 - xvotus$); mapping was competitive across the whole catalogue rather than genome by genome, and no identity or mapping-quality filter was applied at the detection stage, so a read shared between closely related vOTUs is assigned to one of them rather than counted for several, and reads unmapped to the viral pass were then mapped against a prokaryote reference built from the species representatives of the Unified Human Gastrointestinal Genome collection [14], each contig prefixed with its genome accession. The prokaryote pass reproduces the sequential mapping of the published catalogue pipeline so that viral detection is computed on the same read set; no prokaryote-derived quantity enters any analysis reported here. Per-contig statistics were computed with CoverM v0.8.0 [15] in $contig$ mode for the viral pass and

$\geq nome$ mode with $--separa \to r' \mid '$ and $--\ \min\ -covered - \frac{t}{i}on0$ for the prokaryote pass, retaining mean, trimmed mean, covered bases, covered fraction, length and read count; rows with zero coverage were dropped at write time. Detection used a symmetric 50% coverage-breadth threshold on both passes, in contrast to the asymmetric published rule. Runs below three million reads were excluded; the full sample flow is given in Table S4. Biosample overlap with the catalogue is zero.

Catalogue tables were the Unified Human Gut Virome Zenodo deposit downloaded on 2026-07-27 ($read_map\pi ng_relative_abundances$, $read_map\pi ng_samp \leq_m \eta data$, $source_biosamp \leq_m \eta data$, $votus_m\eta data_extended$). The original analyses were reported as running under Python 3.11.15 with numpy 2.4.6 and pandas 3.0.5; rank statistics and their normal-approximation p-values were implemented directly rather than taken from a statistics package. All random number generation used a fixed seed (20260811 for the variant sweep and the delivery-mode and sex-control analyses, 20260812 for the higher-count permutation reruns). The revised primary matching analysis was recomputed from supplied child-mother set counts under Python 3.12.14, numpy 2.3.5 and pandas 3.0.1, using seed 20260909, 2,000 family-label permutations, 10,000 family bootstrap draws and 2,000 one-sample-per-family draws. Portable figure scripts regenerate panels from their supplied numerical source data; this is distinct from repeating upstream read mapping and consensus calling.

Three conventions apply throughout. Presence-absence, prevalence, sharing and identification analyses are at vOTU level; a separate consensus-sequence analysis evaluates within-vOTU consensus-sequence concordance for the subset of genomes with sufficient joint coverage. For the mother-infant cohort the unit of inference is the child, with per-sample values reported for description only. And because the infant-versus-adult prevalence statistic is biased under the null, no raw value of it is compared against one half; every such comparison is referenced to a matched null.

Statistics

The prevalence statistic for a sample is the median, over the vOTUs that sample carries, of the fraction of reference adults carrying that vOTU, with the focal sample's own contribution removed when the focal sample is itself a reference adult. Infant-versus-adult comparison is a rank statistic on those per-sample values.

Matched nulls draw pseudo-focal samples from the reference group, matched on log detection richness and log sequencing depth by nearest neighbour without replacement, and recompute the entire statistic including the leave-one-out step. Three inclusion rules were evaluated: retaining every vOTU present in the adult reference; dropping vOTUs carried by exactly one adult; and restricting to vOTUs carried by at least two adults. All three are reported (Table S2); the main-text values use the second. Calibration used pseudo-infants drawn from adults only, where the true value is one half by construction. This is a secondary analysis of public data and was not preregistered; no formal analysis plan was registered in advance. The primary endpoints are the closed-set identification accuracy and the own-mother sharing

excess; the prevalence comparison against matched nulls and the consensus-concordance contrast are secondary; the delivery-mode, sex-control, retention, temporal-direction and 24-cell sweep analyses are exploratory or sensitivity analyses, and no multiplicity correction is applied across them. Empirical P values were computed in the hypothesised direction of each test: for the infant-prevalence comparisons, where infants are predicted to fall below the matched null, P = (1 + #{null ≤ observed}) / (B + 1); for the sharing, excess and concordance comparisons, where the matched arm is predicted to exceed the null, P = (1 + #{null ≥ observed}) / (B + 1). The smallest attainable value is therefore 1/(B + 1), and a P reported at that value is at the resolution floor rather than below it.

All mother-infant inference is at the child or family level. Pair-level records are used for description only. Family-label permutations shuffle the mother-to-family assignment, preserving every child virome, every maternal detected set and the definition of unrelated sharing. The within-block variant restricts shuffling to families in the same library-number quartile.

Twin sharing within a prevalence stratum is the Jaccard index of both individuals' detected sets restricted to that stratum. The primary mother-child ranking instead uses both full sets; the prevalence-restricted matching sensitivity analysis filters the child set only, as detailed below. Where a contrast has too few units for a permutation null to be informative, an exact permutation over all assignments of the grouping label is used instead and reported as exact; with four families in one arm the attainable P values are discrete and the smallest is bounded away from zero. Every permutation and exact test is one-sided in the direction stated by the hypothesis; a sensitivity calculation using min(1, 2P) gives 0.030 for a one-sided P of 0.015, which remains below 0.05, and 0.056 for the matched-subset ratio P of 0.028, which does not. This doubling calculation is not necessarily identical to a separately defined exact two-sided permutation test. The delivery-mode analyses remain exploratory because the caesarean arm contains only four families, the comparison is confounded by clinical and detection differences, and no multiplicity correction was applied. The statistic differs with the design. Where the units are paired within a child or family — the own-mother-minus-unrelated-mother excess and the maternal-minus-non-maternal retention contrast — the statistic is the median of the paired per-unit differences. Where two disjoint groups are compared, as in the delivery-mode contrast, it is the difference of the group medians. The twin comparison is likewise a difference of medians, between twin pairs and unrelated pairs. The infant-versus-adult comparison in Fig. 4a,b is neither: its statistic is the rank statistic described above, and its null is the matched pseudo-infant draw rather than a label permutation. The exact delivery-mode test enumerates all 35,960 ways of assigning four of the 32 families to the caesarean arm. Paired comparisons across children use the Wilcoxon signed-rank statistic on per-child differences with zero differences dropped and a normal approximation to the null, which is why its P values are not floored in the way the permutation P values are.

Within-vOTU consensus-sequence comparison re-mapped reads to the 309 vOTU representative sequences reaching mean depth 10 or higher in both a child and a mother, identified from the coverage tables of the first pass; the reference was restricted this way because a consensus comparison is uninformative where either member is shallow. Alignment used bowtie2 v2.5.5 with --very-sensitive,

streamed to a sorted alignment without an intermediate file, and consensus bases were called from an mpileup at positions with at least five reads passing a base quality of 20 and a mapping quality of 10, where one base reached 80% of the reads. Positions were compared only where both members had a call. Polymorphic positions were defined per genome as those where at least two distinct bases occur across all individuals carrying that genome, computed before any own-versus-unrelated split, and a pair entered the analysis only with at least 20 shared polymorphic positions. Of the 309 genomes reaching tenfold joint coverage, 241 yielded at least one comparison with 20 or more shared polymorphic positions and entered the 5,011 own and 88,181 unrelated comparisons; of these, 237 had both an own-mother and an unrelated comparison and so appear in the per-genome panel (Fig. S3b), the remaining four contributing only one arm. A mother's timepoints were merged by keeping the deepest call at each position, so a mother's consensus can in principle combine positions called from different timepoints; this favours neither arm, because the same merged mother sequence is used for her own child and for every unrelated child. Polymorphic positions, and the genome and position eligibility rules, were fixed once from the full cohort before any own-versus-unrelated split and were held constant across permutations; only the mother-to-family assignment was permuted. Because the reference is a single representative sequence per vOTU, agreement is measured against that reference frame and residual reference bias or misalignment from closely related genomes cannot be excluded, which is one reason the analysis is described as within-vOTU consensus-sequence concordance rather than strain reconstruction: no haplotype phasing, whole-genome strain assembly, or predefined strain-identity or average-nucleotide-identity threshold was applied. Reads and alignments were deleted as each sample finished; only the per-position consensus calls were retained.

Primary identification ranks all 32 mothers for each child sample by Jaccard overlap of the child's full detected vOTU set and each mother's detected set pooled across her eligible timepoints. No prevalence cutoff enters this primary score. The rank of an assigned mother is the number of candidate scores greater than or equal to her score, including her own score; an all-zero row receives rank 32. All repeated samples from a family receive the same permuted maternal assignment in each null draw. Supplementary prevalence-restricted matching applies an exact cutoff of 3/10 to the child set only; the maternal sets remain unfiltered. This is ≤ 30%, permitting at most 9 of 32 mothers, and is not a one-third cutoff. Leave-out reference panels exclude unique assigned and/or candidate mothers and use their resulting panel size as the denominator (31 or 30, with at most 9 carriers at this cutoff). The unified implementation recomputes assignment-dependent filters for each permuted assignment; completion of that rerun requires the measured per-run profiles described in the code package. Prevalence strata for descriptive sharing are defined within the 32-mother cohort: rare ≤ 10% (0–3 carriers), uncommon > 10% and ≤ 30% (4–9 carriers), and common > 30% (10–32 carriers). They are not catalogue prevalence classes, because only 490 of the 2,843 child-carried vOTUs have a catalogue-adult prevalence. Within each stratum, the own-mother excess is the fraction of the child's stratum-specific set shared with her own mother minus the mean shared with the other 31 mothers, tested against its family-label permutation. Paired comparisons across children use the Wilcoxon signed-rank test on per-child values. Caliper matching, used only in exploratory persistence analysis, pairs each maternal vOTU with a non-

maternal vOTU of the same child whose population prevalence differs by no more than 0.05; unmatched maternal vOTUs are discarded and the retained fraction is reported.

Rarefaction subsamples each detected set to a common size without replacement and repeats the computation. Unless otherwise specified, permutation, rarefaction and one-sample-per-family intervals are the 2.5th to 97.5th percentiles of the corresponding draw distribution; they describe that distribution and are not confidence intervals over families. The 95% confidence interval for mean within-family matching accuracy is instead obtained by resampling the 32 families with replacement and taking the corresponding bootstrap percentiles. Each resampled family contributes one mean accuracy, regardless of its number of samples.

Declarations

Author contributions

Anmin Wang: Methodology, Software, Formal analysis, Visualization, Writing – original draft. Lauren Lai: Validation, Writing – review & editing. Colin Patrick Johnson: Validation, Writing – review & editing. Xiaoyue Yang: Data curation, Writing – review & editing. Chao Ye: Conceptualization, Supervision, Funding acquisition, Writing – review & editing.

Funding

C.Y. discloses support for this work from the National Natural Science Foundation of China (82300707).

Competing interests

The authors declare no competing interests.

Ethics approval and consent

This study is a secondary analysis of de-identified public sequence data and metadata deposited under PRJEB9818 [11] and PRJNA475246 [6], together with the derived Unified Human Gut Virome tables [8]. No new human samples were collected and no identifiable individual data were accessed; ethical approval and participant consent for the original collections are reported in the primary publications.

Data availability

Catalogue tables are available from the Unified Human Gut Virome resource (Zenodo, doi:10.5281/zenodo.17402089), described in the accompanying preprint [8]. Raw sequence data are available from the European Nucleotide Archive and Sequence Read Archive under PRJEB9818 and PRJNA475246. Numerical source files for the figures, analysis-set summary (Table S3), sample flow (Table S4), variant sweep (Table S2) and primary child-mother set-count matrix accompany the supplementary package. The package README distinguishes results reproducible from these derived files from analyses requiring additional per-run coverage profiles and run metadata.

Code availability

Portable primary-matching and figure-generation scripts, dependency specifications, checksums and execution instructions are supplied in the supplementary code package. The primary ranks and their family-label permutation null can be regenerated directly from the supplied set-count matrix. A shared scoring implementation also supports prevalence-restricted and leave-out sensitivity analyses from measured per-run profiles, with assignment-dependent exclusions recalculated under permutation. These profile-dependent analyses require the inputs listed in the README; the supplied summary tables alone do not replace those inputs. The original upstream scripts are retained separately for provenance. No public repository release or DOI is claimed for this package.

References

1. Liang G, et al. The stepwise assembly of the neonatal virome is modulated by breastfeeding. *Nature* 2020;581:470-474. doi:10.1038/s41586-020-2192-1
2. Shah SA, et al. Expanding known viral diversity in the healthy infant gut. *Nat Microbiol* 2023;8:986-998. doi:10.1038/s41564-023-01345-7
3. Lim ES, et al. Early life dynamics of the human gut virome and bacterial microbiome in infants. *Nat Med* 2015;21:1228-34. doi:10.1038/nm.3950
4. Zeng S, et al. A metagenomic catalog of the early-life human gut virome. *Nature Communications* 2024;15:1864. doi:10.1038/s41467-024-45793-z
5. Shamash M, et al. A meta-analysis of infant gut viromes reveals global patterns in bacteriophage community assembly and functional capacity over the first three years of life *Nature Communications* 2026;17(1). doi:10.1038/s41467-026-74609-5
6. Yassour M, et al. Strain-Level Analysis of Mother-to-Child Bacterial Transmission during the First Few Months of Life. *Cell Host Microbe* 2018;24:146-154.e4. doi:10.1016/j.chom.2018.06.007
7. Garmaeva S, et al. Transmission and dynamics of mother-infant gut viruses during pregnancy and early life. *Nat Commun* 2024;15:1945. doi:10.1038/s41467-024-45257-4
8. Camargo AP, et al. A genomic atlas of the human gut virome elucidates genetic factors shaping host interactions. *bioRxiv* 2025. doi:10.1101/2025.11.01.686033
9. Tett A, et al. The Prevotella copri Complex Comprises Four Distinct Clades Underrepresented in Westernized Populations *Cell Host & Microbe* 2019;26(5):666-679.e7. doi:10.1016/j.chom.2019.08.018
10. Carter MM, et al. Ultra-deep sequencing of Hadza hunter-gatherers recovers vanishing gut microbes *Cell* 2023;186(14):3111-3124.e13. doi:10.1016/j.cell.2023.05.046
11. Reyes A, et al. Gut DNA viromes of Malawian twins discordant for severe acute malnutrition. *Proc Natl Acad Sci U S A* 2015;112:11941-6. doi:10.1073/pnas.1514285112
12. Maqsood R, et al. Discordant transmission of bacteria and viruses from mothers to babies at birth. *Microbiome* 2019;7:156. doi:10.1186/s40168-019-0766-7

13. Langmead B, et al. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 2012;9:357-9. doi:10.1038/nmeth.1923

14. Almeida A, et al. A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat Biotechnol* 2021;39:105-114. doi:10.1038/s41587-020-0603-3

15. Aroney STN, et al. CoverM: read alignment statistics for metagenomics. *Bioinformatics* 2025;41(4):btaf147. doi:10.1093/bioinformatics/btaf147

Figures

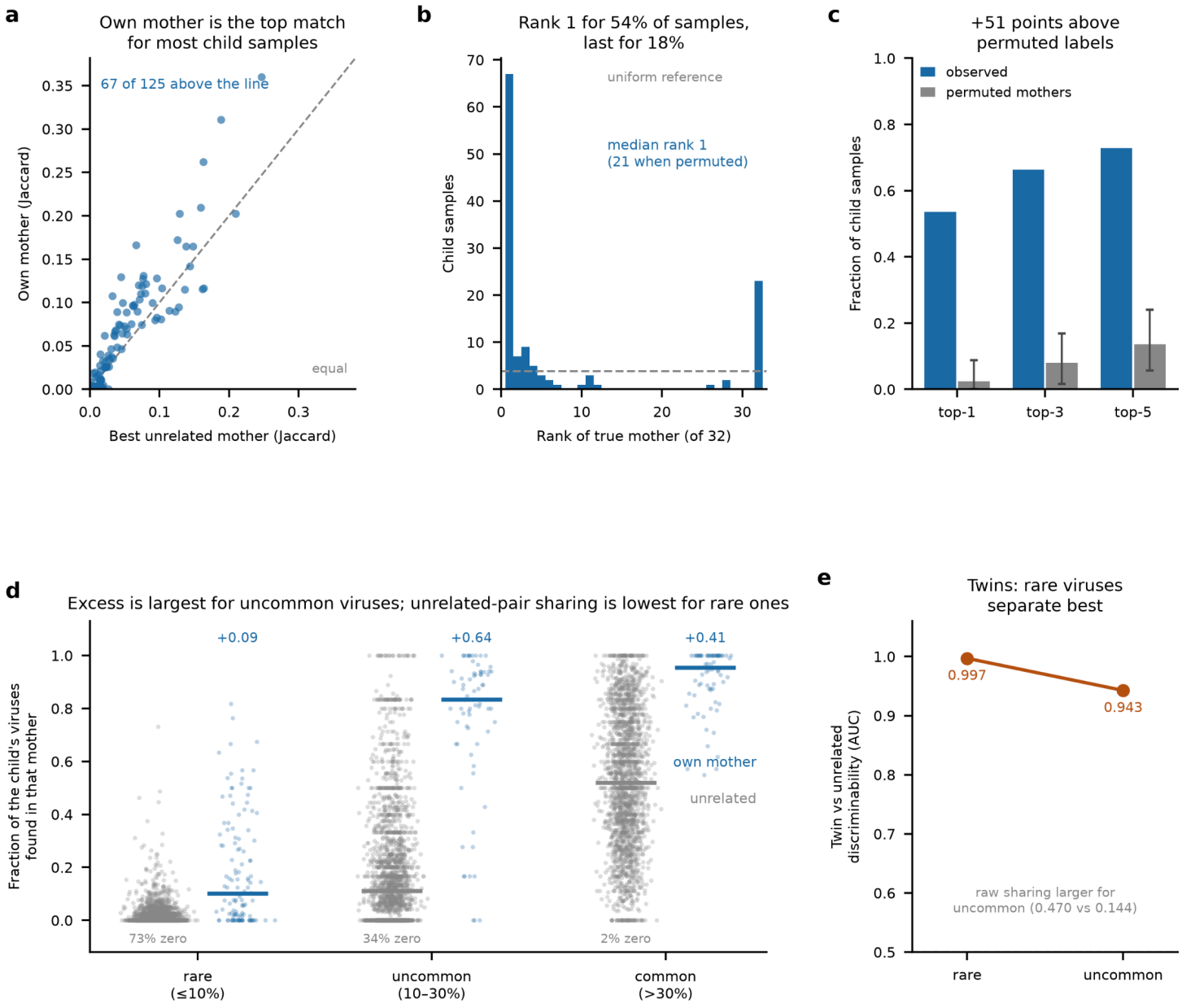


Figure 1

Closed-set matching of a child's virome to their own mother. (a) For each of 125 child samples, Jaccard overlap between the child's full detected vOTU set and the child's own mother (y) against the highest-scoring unrelated mother (x); points above the dashed line are samples whose own mother is the top match. (b) Distribution of the rank of the true mother among all 32 mothers, one observation per child sample; the dashed line is the uniform expectation of 125/32 samples per rank. Ties are counted against the true mother, so a sample sharing nothing with any mother is assigned the last rank. No maternal-prevalence filter enters panels a–c. (c) Top-one, top-three and top-five accuracy (blue) against the same quantity after permuting which mother belongs to which family (grey, median with 2.5th-97.5th percentile range over 2000 permutations), preserving the observed viromes, ties and repeated samples within families. (d) Fraction of a child's detected viruses that are also found in a given mother, split by how many of the 32 mothers carry the virus; each point is one child-mother pair, bars are medians, the blue number above each stratum is the own-mother excess from a family-label permutation and the grey number below is the share of unrelated pairs sharing nothing. (e) In the twin cohort, the area under the ROC curve separating co-twin pairs from unrelated pairs, by prevalence stratum; the annotation gives the raw median sharing for comparison, which runs the other way.

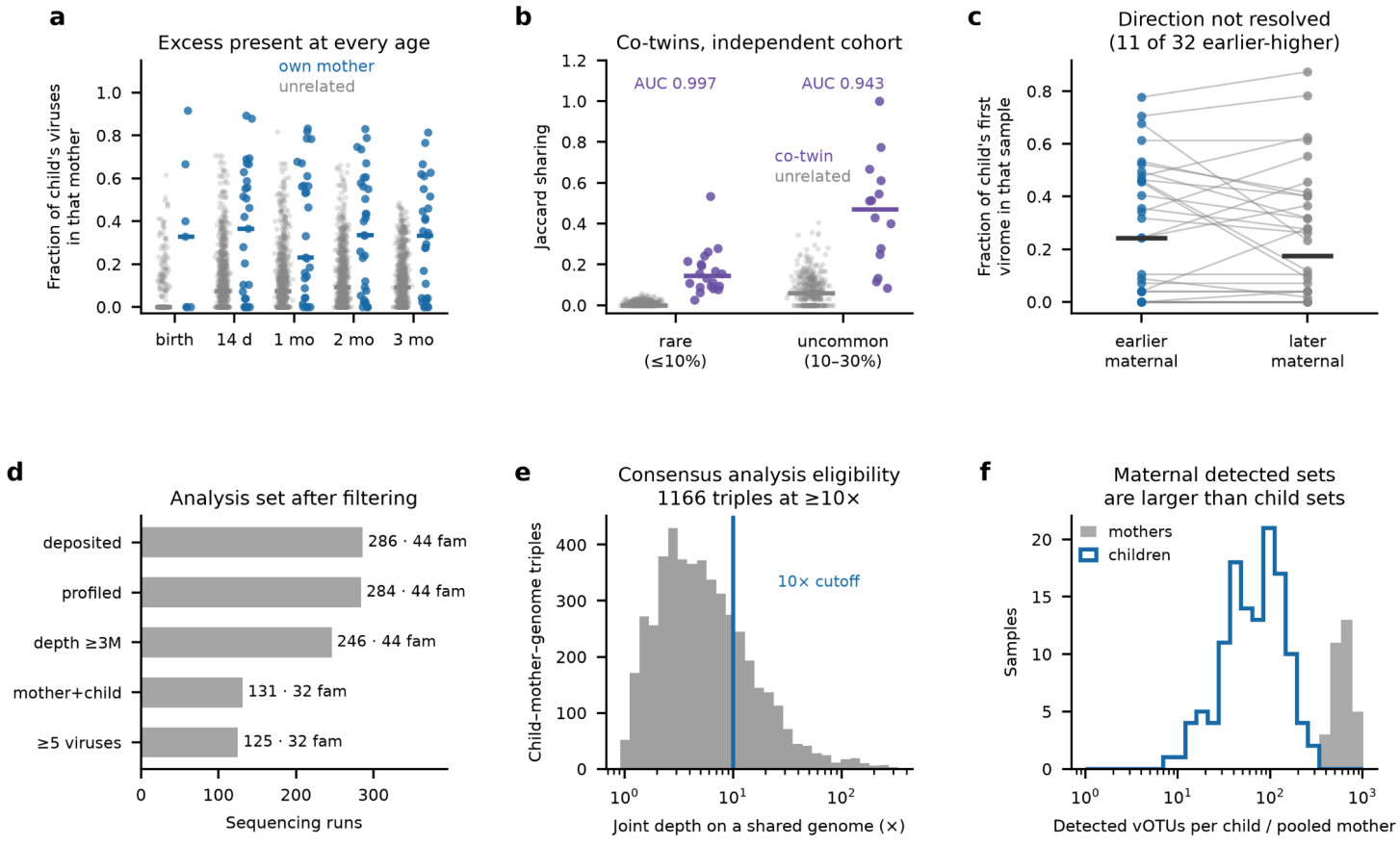


Figure 2

Cohort structure, sampling, and what bounds each analysis. (a) Fraction of a child's detected viruses also found in a given mother, split by the child's age, with each child-mother pair drawn as a point and medians as bars; the descriptive own-mother excess is positive at every sampled age (+0.33 at birth,

+0.29 at 14 days, +0.16 at one month, +0.24 at two months, +0.24 at three months). The birth column rests on 7 children because most meconium samples fell below the depth threshold. (b) The independent twin cohort: Jaccard sharing for co-twin and unrelated pairs by prevalence stratum, with the discriminability above each pair of columns; raw sharing is higher for uncommon viruses while separation is near-perfect for rare ones. (c) Fraction of each child's first post-birth virome found in an earlier versus a later maternal sample, one line per family; earlier maternal samples do not predict the child better, so these data do not resolve direction. (d) Runs and families surviving each filter, ending at the set used for the sharing statistics. (e) Joint sequencing depth for every child-mother-genome triple sharing a virus; the consensus-concordance comparison uses the 1,166 triples at or above tenfold, which bounds what it can resolve. (f) Detected vOTUs per pooled maternal set and per child sample (medians 601 and 69). These are different aggregation units; the maternal set pools eligible timepoints, whereas each child observation is one sample.

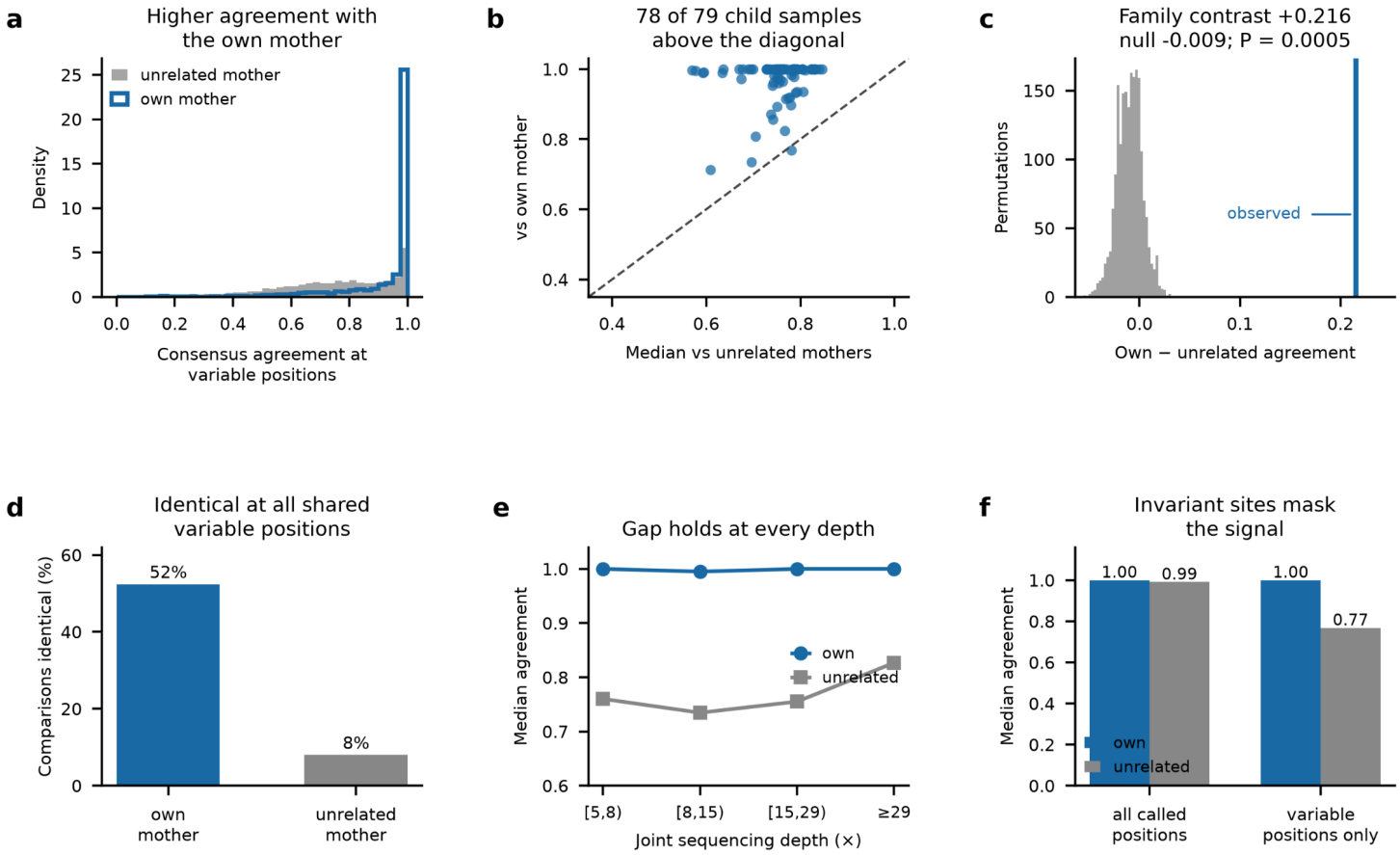


Figure 3

Shared viruses show greater within-vOTU consensus-sequence concordance in matched mother-child pairs. Reads were re-mapped to the 309 viral genomes reaching tenfold coverage in both a child and a mother; a consensus base was called at each position covered at least fivefold with at least 80% agreement, and only positions called in both members of a pair were compared. (a) Distribution of consensus agreement at variable positions for own-mother (outline) and unrelated-mother (filled)

comparisons; 5,011 own and 88,181 unrelated comparisons across 241 genomes. (b) One point per child sample: median agreement with the own mother against the median across unrelated mothers; the dashed line is equality. Samples repeat within children, so the test is performed at the family level. (c) The family-level contrast (+0.216) against its family-label permutation null (median −0.009), evaluated on 24 families (B = 2,000; P = 1/2,001). (d) Fraction of comparisons identical at every shared variable position. (e) Within-quartile median agreement by joint sequencing depth, shown without uncertainty; greater own-mother concordance is present in every quartile, although this descriptive stratification does not exclude all depth-related bias. (f) Why the restriction matters: across all called positions an unrelated mother already matches at 0.99 because only 2.6% of called positions vary across the cohort (a median of 190 of 9436 per genome), and the arms separate only on those variable positions.

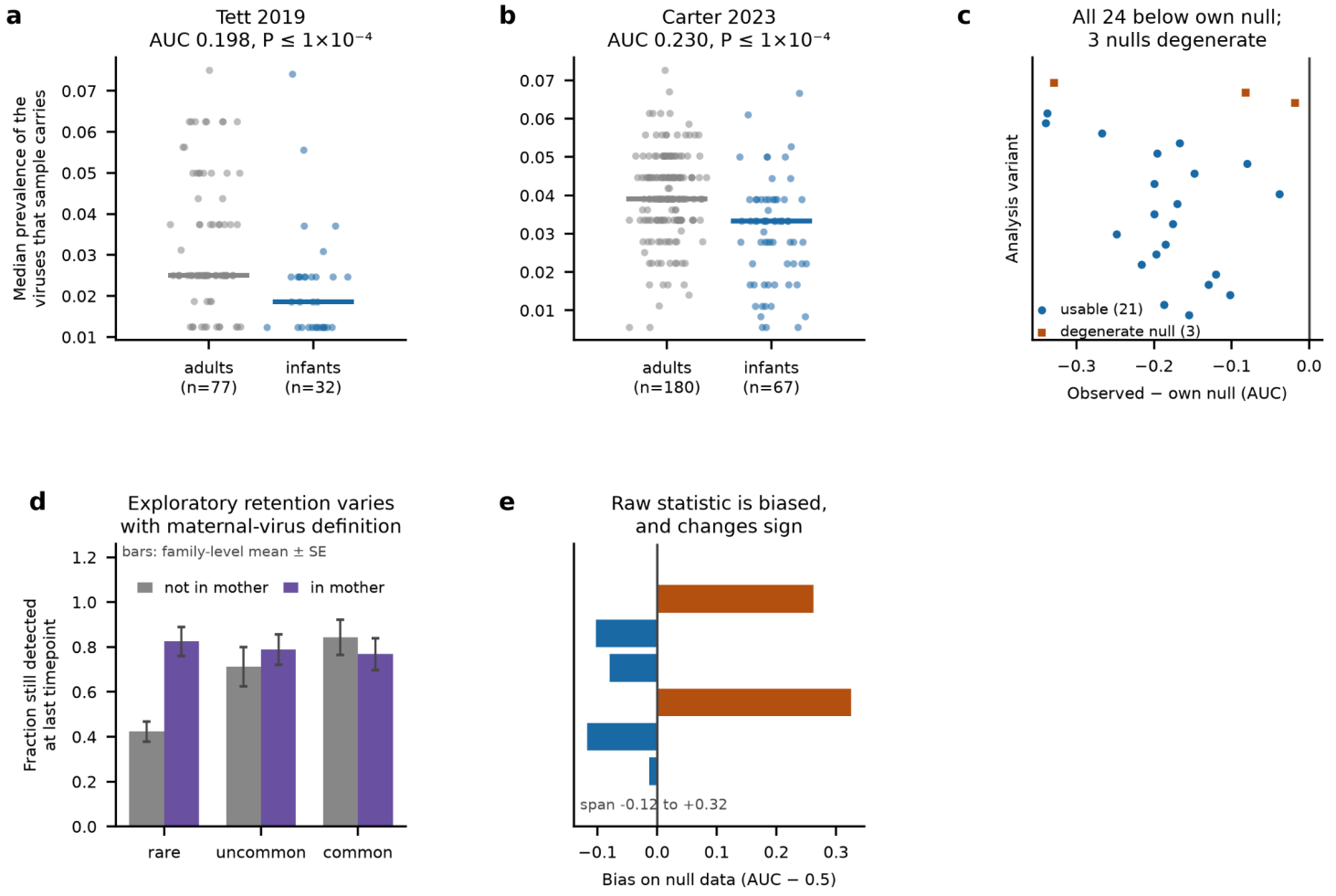


Figure 4

The premise, its calibration, and an exploratory retention comparison. (a) For the Tett 2019 study and (b) the Carter 2023 study, the median adult-population prevalence of the viruses each sample carries, one point per sample, with medians as bars; titles give the rank statistic and its empirical P against a richness- and depth-matched pseudo-infant null at 10,000 permutations. (c) Each of the 24 analysis

variants as its observed value minus that variant's own null; all 24 lie below their null, and the three variants whose null collapsed to a single value are marked and excluded from the summary claim. (d) Fraction of viruses present at a child's first post-birth timepoint still detected at the last, by prevalence stratum and by whether the mother also carried the virus. Bars are means over all families contributing to each arm, with standard errors computed across families rather than across vOTUs, since vOTUs cluster within children; the paired comparison quoted in the text restricts to the 19 families contributing to both arms, giving a slightly different non-maternal mean (45.2% against 42.3% here); the stratum-by-maternal-origin interaction was not supported (permutation P = 0.66) and the comparison is reported as exploratory. (e) The same prevalence statistic computed on data constructed to contain no age effect, where the true value is one half; bars give the bias, which spans −0.12 to +0.32 and changes sign with the inclusion rule.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- infantviromesupplement.docx
- tableS4sampleflow.csv
- tableS1circularity.csv
- tableS3analysissets.csv
- tableS2q1variantsweep.csv