bioRxiv preprint doi: https://doi.org/10.1101/460113; this version posted October 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Acquisition, transmission and strain diversity of human gut-colonizing crAss-like phages

Benjamin A. Siranosian[1], Fiona B. Tamburini[1], Gavin Sherlock[1], Ami S. Bhatt[1,2]*

[1]Department of Genetics, Stanford University, Stanford, CA, USA.

[2]Department of Medicine, Division of Hematology, Stanford University, Stanford, CA, USA.

*To whom correspondence may be addressed.

Abstract

CrAss-like phages are double-stranded DNA viruses that are prevalent in human gut microbiomes. Here, we analyze gut metagenomic data from mother-infant pairs and patients undergoing fecal microbiota transplantation to evaluate the patterns of acquisition, transmission and strain diversity of crAss-like phages. We find that crAss-like phages are rarely detected at birth but are increasingly prevalent in the infant microbiome after one month of life. We observe nearly identical genomes in 50% of cases where the same crAss-like clade is detected in both the mother and the infant, suggesting vertical transmission. In cases of putative transmission of prototypical crAssphage (p-crAssphage), we find that a subset of strains present in the mother are detected in the infant, and that strain diversity in infants increases with time. Strain diversity in the p-crAssphage population in mothers is generally low, with a median of 50 variable sites in the 97kb genome. Putative tail fiber proteins are enriched for nonsynonymous strain variation compared to other genes, suggesting a potential evolutionary benefit to maintaining strain diversity in specific genes. Finally, we show that p-crAssphage can be acquired through fecal microbiota transplantation. These results enhance our understanding of the acquisition, individual-level strain diversity, and transmission of crAss-like phages, which have emerged as some of the most successful human-associated microbes.

bioRxiv preprint doi: https://doi.org/10.1101/460113; this version posted October 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Main text

Introduction

In addition to trillions of bacteria, the human gastrointestinal tract is densely populated with bacteriophages. Bacteriophages can drive bacterial community composition and mediate horizontal gene transfer[1], and alterations in the human gut virome have been associated with disease[2,3]. However, our knowledge of the contributions of specific bacteriophages to human biology is limited, in part due to the paucity of viral sequences represented in reference databases. High-throughput sequencing and advanced genomic tools have facilitated the *in silico* discovery and characterization of previously unknown bacteriophages. The preeminent example of such a discovery is crAssphage (**cr**oss-**Ass**embly phage), initially identified from human virome sequencing data[4]. A bacteriophage with a ~97 kilobase circular, double-stranded DNA genome, crAssphage sequences are found almost exclusively in human fecal metagenomes in diverse populations globally[5–8], and can be highly abundant. Initial estimates indicate that crAssphage is present in up to 73-77% of humans[4,5]. Given the near ubiquity of crAssphage and its apparent specificity to the human gut, quantitative PCR assays have been developed to use crAssphage genes as markers for tracking human fecal pollution in water and environmental samples[9,10] and in human stool[11].

More recent investigations have shown that crAssphage is one member of a wide range of crAss-like phages that exist in the human microbiome[5,12]. In this manuscript, we adopt the taxonomic classification system for crAss-like phages by Guerin et al. [5], which proposed 4 subfamily (Alpha, Beta, Gamma, Delta) and 10 cluster (1-10) designations based on shared protein coding genes. The crAssphage first described by Dutilh et al.[4] belongs to the Alpha subfamily, cluster 1 and is given the designation prototypical crAssphage (p-crAssphage); we use p-crAssphage in all further designations to avoid ambiguity. The genomes classified as

bioRxiv preprint doi: https://doi.org/10.1101/460113; this version posted October 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

crAss-like phages by Guerin et al. are diverse - members of the same cluster share at least 40% of protein coding genes, while members of the same subfamily share only 20%-40% of protein coding genes.

It is not known whether or how crAss-like phages influence host biology or disease [13,14]. To answer higher-order questions about the role of crAss-like phages in human biology, it is necessary to establish basic principles of acquisition, persistence, and distribution. While p-crAssphage has been detected in infant gut metagenomes[7,13], it is not yet known how crAss-like phages are acquired in infancy. Infants acquire many of their first microbes, such as *Bacteroides* species, from their mother during and after delivery[15–18]. By contrast, it has been demonstrated that adult twins and their mothers have unique gut viromes [19]. Given that *Bacteroides* species are hypothesized to be the bacterial host(s) of p-crAssphage[4,20], and the apparent specificity of p-crAssphage to the human gut as opposed to other mammals or environmental samples, we postulated that p-crAssphage is vertically transmitted from mother to infant, similar to what is observed for many bacterial taxa and in contrast to what is reported for other members of the human virome. To test this hypothesis, we examined publicly available shotgun metagenomic data from two stool microbiome datasets[15,16] consisting of samples from mothers and their infants (n=143 mother-infant pairs).

In this study, we find that p-crAssphage and other crAss-like phages are rarely detected in the microbiome at birth but become detectable during the first year of life. We observe >99.5% identical genome sequences in about one quarter of cases where mothers and infants have the same crAss-like phage, suggesting vertical transmission from mother to infant. Infants acquire a reduced diversity population of p-crAssphage compared to their mother, but strain diversity expands upon colonization. Finally, by examining shotgun metagenomic data from patients undergoing fecal microbiota transplantation (FMT), we show that FMT recipients can acquire p-crAssphage with a nearly identical genome sequence as the stool donor. These results begin to uncover the principles of acquisition and transmission of p-crAssphage and

bioRxiv preprint doi: https://doi.org/10.1101/460113; this version posted October 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

other crAss-like phages, which are the most prevalent human-associated phages described, to date.

Results

We evaluated the presence and abundance of p-crAssphage in the microbiomes of mothers and infants by classifying sequencing reads with Kraken2[21], using a database of all bacterial, viral and fungal genomes in NCBI GenBank assembled to complete genome, chromosome or scaffold quality as of February 2019 (see Methods). P-crAssphage is represented by a 97kb genome (accession NC_024711.1). Assigning absolute presence or absence of an organism in metagenomic sequencing data is difficult and confounded by sequencing depth. Here, we consider samples with ≥1,000 reads classified as p-crAssphage to be evidence for presence, as this corresponds to roughly 1x coverage of the genome (assuming 100bp reads and a ~100kb genome length). Samples from mothers and infants had an average of 8.7M reads after preprocessing, and the 1,000 read coverage threshold thus corresponds to an average relative abundance of 0.011%. Of note, this somewhat arbitrary threshold, while fairly specific, renders our approach limited in sensitivity - that is, we do not report on p-crAssphage when it is present at lower relative abundance.

Although p-crAssphage is highly abundant in the adult gut microbiome[5,14] and has been detected in infant gut microbiomes, it is unclear when or how it is acquired. Consistent with previous studies describing low relative abundance or absence of p-crAssphage in the infant gut microbiome[7,22], we found 0 out of 22 infants have ≥1,000 p-crAssphage reads in samples collected within 24 hours of birth (Table S1). P-crAssphage increases in prevalence as infants age: it is detected in 3/35 (9%) infants from Yassour et al.[16] by three months and 16/100 (16%) infants in Bäckhed et al.[15] by 12 months. P-crAssphage is more prevalent in adult mothers, where it is detected in at least one sample from 8/35 (23%) and 25/100 (25%) of mothers in

bioRxiv preprint doi: https://doi.org/10.1101/460113; this version posted October 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

each study, respectively (Figure S1). P-crAssphage is detected in both the mother and her infant in ten cases, while 23/33 p-crAssphage positive mothers have a p-crAssphage negative infant, and 9/19 p-crAssphage positive infants have a p-crAssphage negative mother.

The infant gut microbiome is strongly impacted by delivery mode, and it has been shown that infants born by Cesarean section initially lack *Bacteroides* species[16,18,23–25]. In line with the described effects of Cesarean section delivery and a hypothesized Bacteroidetes host, we found that all 19 p-crAssphage positive infants were delivered vaginally, while all 21 infants delivered through Cesarean section remained p-crAssphage negative. Although not significant when each study is tested individually, the association between delivery mode and p-crAssphage presence is significant when samples from the two studies are combined (p=0.043, Fisher's exact test). This result contrasts with the findings of McCann et al.[7], where the authors found no association between p-crAssphage relative abundance and delivery mode.

Vertical transmission of gut microbes from mother to infant is common and well-described among certain bacterial taxa[15–17,26]. To test the hypothesis that p-crAssphage can be vertically transmitted from mother to infant, we investigated metagenome-assembled p-crAssphage genomes from ten p-crAssphage positive mother-infant pairs (see Methods). In six cases, mother-infant pairs had nearly identical assembled sequences (families M0226, M0808, M1098, 335, 343 and 345; >99.7% similarity). Two mother-infant pairs had assembled sequences more similar than unrelated pairs (families 263 and 268; 98-99.2% similarity), and two mother-infant pairs had assembled sequences that were no more similar than unrelated pairs (families 184 and 272; ~96% similarity) (Figure 1). Overall, related mothers and infants harbor more closely related p-crAssphage sequences than unrelated mothers and infants (Figure S2a). When all samples with sufficient p-crAssphage coverage were included in the assembly comparison, no pairs from unrelated individuals had > 98% similarity (Figure S3). Assembled p-crAssphage genomes were high quality and contiguous: in the 29 samples from families with p-crAssphage found in mothers and infants, the median contig N50 was 59.6kb

bioRxiv preprint doi: https://doi.org/10.1101/460113; this version posted October 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(standard deviation, SD= 41.4kb), median number of contigs was 3 (SD=17) and median total assembled length was 96.3kb (SD=24.2kb) (assembly statistics reported in Table S2). One-to-one pairwise alignments between assembled sequences from mother-infant pairs had a median length of 85.8kb (SD=27.8kb). 22 samples assembled a nearly complete (>95kb) p-crAssphage genome in a single contig. The assembled genomes also share 91.2% - 97.6% nucleotide identity with the p-crAssphage reference genome, adding confidence that they are truly representative assemblies. Next, we used a variant calling approach to identify fixed SNPs compared to the p-crAssphage reference (see Methods). Pairs of samples were compared at genomic sites covered ≥10x and used to construct a heatmap of SNP similarity. The same six mother-infant pairs (M0226, M0808, M1098, 335, 343 and 345) had >99.5% SNP similarity and continued to cluster together (Figure S4). Mother-infant pairs had higher SNP identity than unrelated pairs on average (Figure S2b).

Metagenomic assembly only represents the dominant allele at each position, and a fixed SNP comparison only considers sites that are identical across all strains present in a sample. To understand the differing p-crAssphage strains present in the microbiome, we would ideally phase strain “haplotypes” with a technology like long-read sequencing. With only short reads available, we examined genomic positions that had multiple single nucleotide variant alleles called at high-quality (≥5 reads for each allele, multiallelic sites) as a proxy for strain diversity. We report a normalized statistic (F_{multi}) to compare multiallelic sites across samples with highly variable coverage. At a given minor allele fraction(AF), F_{multi} is the proportion of multiallelic sites with a minor AF > *x* among those sites covered well enough to detect a minor AF of *x.*

We calculated F_{multi} for minor allele fractions of 0.40, 0.30, 0.20, and 0.10 and compared across samples at a given minor AF value. At every AF tested, infants as a group had a smaller fraction of multiallelic sites when compared to mothers as a group. (Figure 2a). In the three cases where we detected p-crAssphage in multiple samples from the same infant, we found more multiallelic sites in later samples. We were typically powered enough to detect multiallelic

bioRxiv preprint doi: https://doi.org/10.1101/460113; this version posted October 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

sites of the observed AF in earlier samples, but we cannot rule out the possibility that newly observed variants are below our limit of detection in earlier samples. Multiallelic sites in infants are often fixed sites in the p-crAssphage population of the mother (Figure 2d). In contrast to infants, mothers from Yassour et al.[16] showed no change in the proportion of multiallelic sites over the 6 month sampling period (Figure 2b). We then looked at multiallelic sites in mothers that are fixed in matched infant samples. In 2/3 cases where we observed putative p-crAssphage transmission and the mother had ≥10 multiallelic sites, major alleles are disproportionately detected in the child (Figure 2c). Taken together, these results suggest a few potential models: one is that infants acquire a single strain or limited diversity of p-crAssphage strains. As the infant microbiome matures and diversifies with age, the p-crAssphage population can evolve and acquire genetic diversity. Alternatively, a larger spectrum of p-crAssphage strains than is detected may be harbored in a mother, with some strains below the limit of our detection. These strains may experience less selective pressure in the infant than in the mother; thus, these strains may be more numerous and easily detected in the infant than they are in the mother.

We continued to use multiallelic sites to investigate p-crAssphage strain diversity within an individual. Mothers generally have limited strain diversity, with a median of 0.41 (SD=7.5) multiallelic sites per kb at AF > 0.1; variation in 0.04% of the genome. In a small number of cases, we do observe samples with up to 100x more frequent multiallelic sites (Figure 3a). This allelic variation could be the result of many closely related strains existing together or a smaller number of more divergent strains. Phasing strain “haplotypes” is necessary to distinguish between these possibilities. Limited strain diversity suggest an exclusion principle, which favors mono- or oligo-colonization of a particular p-crAssphage strain or closely-related strains within the gut of an individual, though notably, a minority of individuals may be simultaneously colonized by multiple diverse strains. P-crAssphage has fewer multiallelic sites on average compared to Lactococcus phages, which are the only other group of phages detected at ≥1x

bioRxiv preprint doi: https://doi.org/10.1101/460113; this version posted October 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

coverage in at least ten samples. Using the reference genome of the most frequently detected individual phage, *Lactococcus phage 16802,* 34 samples had at least 1x coverage; these samples had a median of 58.8 (SD=14.7) multiallelic sites per kb at AF > 0.1 (Figure S9).

We next evaluated whether strain variation in the p-crAssphage population was the result of synonymous or nonsynonymous genomic changes. Variant effects were predicted using the p-crAssphage genome annotation from GenBank and SnpEff[27] (see Methods). We compared the proportion of observed variant effects to a null model of equal probability of mutation at every base in the reference genome. In samples from mothers, multiallelic sites with predicted nonsynonymous and nonsense effects were less likely than expected under the null model, while synonymous sites were more likely than expected ($p < 1e-5$ for each category, likelihood ratio test). (Figure 3b). An overrepresentation of synonymous variants suggests that strain diversity in the p-crAssphage population of mothers is enriched for neutral genetic variation, which may have been acquired over the relatively long time the phages could have been present in the microbiome. In contrast to mothers, the predicted effects of multiallelic sites in infants were indistinguishable from the null model ($p > 0.05$ for each effect category, likelihood ratio test). This may suggest that multiallelic sites in infants arise randomly and the forces acting to influence the distribution of predicted sites in mothers have not had time to act on the infant's p-crAssphage population yet. Alternatively, selective pressures acting on p-crAssphage alleles may be entirely different in the infant and mother microbiome. Comparing mother and infant distributions showed that only the proportion of synonymous multiallelic sites was significantly different between the two ($p = 0.04$, likelihood ratio test). We note that synonymous variants may not be truly neutral, as noncoding variants have been shown to affect translation efficiency in bacteriophages[28].

In adults, we find wide variation in the number of multiallelic sites across the p-crAssphage genome, with enrichment in the number of sites and the ratio of nonsynonymous to synonymous variants corresponding to certain predicted genes (Figure 4). Multiallelic sites were

bioRxiv preprint doi: https://doi.org/10.1101/460113; this version posted October 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

detected in 80/88 predicted genes. When genes were ranked by the length-normalized number of nonsynonymous variants, "putative Tail sheath protein" was the top annotated gene, and other predicted tail proteins also had high ratios (Table S4). Phage tail proteins are responsible for host tropism[29–31]; therefore maintaining multiple functionally different alleles in the population may be beneficial to expand the host range of p-crAssphage. Increased variation in tail fiber genes was also found in an analysis of p-crAssphage genomes from South Africa[32]. The genes that were least likely to have nonsynonymous multiallelic sites appear to be those that are critical for phage function, such as "putative portal protein", "putative major capsid protein" and putative RNA polymerase subunits. Interestingly, some of these genes correspond to peaks in the number of multiallelic sites detected, even though the variants had mostly synonymous predicted effects. Infant samples have fewer multiallelic sites than mothers, with sites detected in 33/88 genes. "Putative ssb single stranded DNA-binding protein" was the most enriched gene for nonsynonymous multiallelic sites. Five out of ten tail fiber proteins had at least one nonsynonymous multiallelic variant in infant samples, and the most frequently mutated tail fiber gene was "putative phage tail-collar fiber protein (DUF3751)," a top hit in the adult samples (Table S4).

P-crAssphage is the first described member of an expanding group of crAss-like phages. Guerin et al.[5] assembled 249 complete or near-complete crAss-like phage genomes out of metagenomic sequencing datasets, which were then classified into 4 subfamilies (Alpha, Beta, Gamma, Delta) and 10 clusters (1-10) based on shared protein coding genes. P-crAssphage is a member of cluster Alpha 01. Given the observed sharing of p-crAssphage genome sequences by mother-infant pairs, we were interested to determine if similar putative transmission events could also be observed for crAss-like phages. We added the crAss-like genomes from Guerin et al.[5] to the Kraken2 viral reference database in a hierarchy following the proposed subfamily and cluster designations (see Methods). For classification and transmission analyses, we carried out

bioRxiv preprint doi: https://doi.org/10.1101/460113; this version posted October 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

analyses at the level of crAss-like phage clusters. A threshold of 1,000 reads classified to the same cluster (roughy 1x coverage) was treated as evidence for presence.

Broadly, crAss-like phages are more frequently detected in the microbiome of mothers and infants than p-crAssphage alone. 7/36 (19%) infants from Yassour et al.[16] and 49/100 (49%) infants from Bäckhed et al.[15] have at least one crAss-like cluster detected in at least one sample. At least one cluster was detected in 33/43 (77%) and 88/100 (88%) of mothers from each study. Mothers are most likely to be colonized by a single crAss-like phage cluster, although we observe samples with up to 8 clusters detected (Figure S5). We do not observe any crAss-like phage cluster present in infant samples taken within 24 hours of birth. However, two infants from Bäckhed et al.[15] have a crAss-like phage meeting the presence threshold in samples collected as soon as 3 days after birth (samples 385_B and 633_B). This may represent the lower time limit for the crAss-like phage and its bacterial hosts(s) to reach the detection threshold.

The putative hosts of crAss-like phages, members of the Bacteroidetes phylum, are known to be vertically transmitted[15–17,26]. Thus, we hypothesized that crAss-like phages would be more frequently transmitted to vaginally vs. C-section born infants. We observed that all 19 p-crAssphage positive infants were delivered vaginally, while all C-section born infants remained p-crAssphage negative for the duration of sampling. Although low sample numbers prevented this association from raising to the level of significance when each cohort was tested individually, it was significant when samples from both cohorts were considered together (Bäckhed p=0.12, Yassour p=1, combined p=0.043). We also tested for associations in cases where at least 10 infants were positive for a given crAss-like phage cluster. Presence of cluster Delta 07 (Bäckhed p=0.009, Yassour p=1, combined p=0.007) and *any* crAss-like phage cluster (Bäckhed p=0.004, Yassour p=0.32, combined p=0.001) was significantly associated with vaginal delivery (Table S1). P-values were calculated with Fisher's exact test and are

bioRxiv preprint doi: https://doi.org/10.1101/460113; this version posted October 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

uncorrected for multiple hypothesis testing. No significant associations between crAss-like phage presence and breastfeeding status were found.

Although a crAss-like phage similar to cluster Beta 06 phages was recently cultured on a *Bacteroides intestinalis* host[20], the hosts of other crAss-like phages have yet to be identified. We searched for bacterial taxa that were differentially abundant between crAss-like phage positive and negative infants to make inferences about potential hosts. Samples from vaginally born infants at three or four months of age were included to allow comparisons across the two studies at a similar time point. Bacterial relative abundances were transformed to centered log-ratios, and differential abundance was calculated with the R package ALDEx2[33] (see Methods). P-values were calculated with the two-sided Wilcoxon rank-sum test and corrected for multiple hypothesis testing[34]. Due to limited sample numbers, we considered presence of *any* crAss-like phage as a group.

At the genus level, Collinsella was the most enriched taxon in infants positive for any crAss-like phage (corrected p=0.0167) (Figure S6, Table S6). Several members of the Collinsella genus, including *Collinsella aerofaciens* (corrected p=0.0239) were also the most enriched species in these infants. Certain species of the genus Bacteroides were also significantly enriched to a lesser degree, such as *Bacteroides massiliensis* (corrected p=0.0499). Collinsella is a member of Actinobaceria, an entirely different phylum than the posited Bacteroidetes hosts[4]. Collinsella was previously identified as a signature of the developing anaerobic infant microbiome[15], but further work is necessary to determine if these species have a direct or indirect influence on the acquisition of crAss-like phages.

Next, we searched for putative mother-infant transmission of crAss-like phages. Depending on the cluster, 0-16 families (median 2.5, SD=5.1) have the same crAss-like phage cluster detected in at least one mother and matched infant sample (Table S5). We extended the metagenomic assembly analysis above, using a pangenome compiled from all crAss-like genomes in each cluster in place of the p-crAssphage reference. We find cases where mothers

bioRxiv preprint doi: https://doi.org/10.1101/460113; this version posted October 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

and infants share an assembled genome with >99.7% identity and at least 20kb of aligned sequence in 7/10 clusters. As expected, candidate transmission events in cluster Alpha 01 matched p-crAssphage. Overall, we observe putative transmission in 21/42 (50%) cases where mother and infant have the same crAss-like phage cluster (Table S5). In two families, we observed putative transmission of two separate crAss-like phage clusters. 50% is likely an underestimate of the true transmission rate, because many comparisons were limited by low sequencing depth and poorly assembled draft genomes. Interestingly, a subset of infants have a crAss-like phage but their mothers do not. This could be due to waxing/waning amounts of crAss-like phages in mothers, as has been described in adults[14]. If this is the case, the crAss-like phage may have been present in the mother at a level lower than our limit of detection and thus may have been transmitted to the baby. Alternatively, the baby may have acquired the crAss-like phage from an altogether different source, such as another housemate or environmental source.

Another example of a perturbation where the gut microbiome, typically stable in adults, may acquire a large number of new microbes is when individuals experience infection with the gut pathogen *Clostridium difficile,* are treated with antibiotics, and subsequently receive fecal microbiota transplantation (FMT)[35,36]. Draper et al.[37] found that p-crAssphage relative abundance is decreased in individuals with recurrent *C. difficile* infection and that p-crAssphage could be transplanted from donor to recipient. However, strain-level transmission of p-crAssphage has not been explored in this patient population. We examined metagenomic sequencing data from Smillie et al.[35] and viral metagenomic sequencing data from Draper et al.[37] using the classification, assembly and comparison methods described above.

In the data from Smillie et al.[35], we detect p-crAssphage at ≥1x coverage in samples from two donors, MGH06D and MGH03D (Table S1). 12 patients received stool preparations from either of those two donors. After FMT, 8/11 (73%) patients who received material from donor MGH03D were positive for p-crAssphage, while the individual who received material from

bioRxiv preprint doi: https://doi.org/10.1101/460113; this version posted October 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

donor MGH06D remained negative (Figure 5a). Zero recipients who received FMT from a p-crAssphage-negative donor acquired p-crAssphage during the sampling period. We compared assembled p-crAssphage genomes from donors and recipients and found > 99.8% identical sequences in samples from MGH03D and recipients of this donor's material, while samples from donor MGH06D had a distinct p-crAssphage sequence (96.7% identity to MGH03D) (Figure S7). Interestingly, individuals MGH11R and MGH12R experienced dynamic p-crAssphage presence, with the phage falling below and rising above the detection limit in subsequent samples. The assembled genomes remained highly similar in each case, suggesting a waxing/waning p-crAssphage population in individual MGH12R, who did not receive an additional FMT. In pre-FMT samples from recipients from this study, p-crAssphage was not detected, and only one sample was positive for a single crAss-like phage (Table S1). This suggests the phages are substantially diminished in abundance when individuals are treated with drugs such as metronidazole, which has high activity against *Bacteroides species*.

Draper et al.[37] specifically sequenced the amplified viral content of the metagenome, so we adjusted the detection threshold to 10,000 reads (10x coverage) to reduce the number of false positives. P-crAssphage was detected in all 16 samples from donor D3 and 0/17 samples from donors D1 and D2. No pre-FMT samples from *C. difficile* colitis affected patients were positive for p-crAssphage or any crAss-like phage (Table S1). 7/7 patients who received FMT from donor D3 material became p-crAssphage positive; most remained positive for the 12 month duration of sampling (Figure 5b). Assembled p-crAssphage genomes from donor D3 and the seven recipients had >99.5% nucleotide identity, suggesting colonization with the specific donor p-crAssphage strain (Figure S8). Patient P7 became p-crAssphage positive with a genome ~92% identical to the other donor and patients. P7 received material from donor D1, who was p-crAssphage negative, and therefore could have acquired the phage from a population below the detection limit in the donor or another source following re-establishment of host bacterial populations. Patient P13 had p-crAssphage present at ~40x in a single sample,

bioRxiv preprint doi: https://doi.org/10.1101/460113; this version posted October 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

but the assembled genome only had a total length of 12kb and N50 of 2.8kb. Samples with lower coverage assembled nearly complete p-crAssphage genomes with N50 >60kb. It is thus possible that P13's p-crAssphage detection is an artifact of PCR amplification that is often used in the sequencing of virus-enriched samples. These data show that p-crAssphage is frequently and efficiently transplanted via FMT and that p-crAssphage can stably engraft in FMT recipients for up to one year.

bioRxiv preprint doi: https://doi.org/10.1101/460113; this version posted October 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Discussion

The *in silico* discovery of p-crAssphage and recent publication of hundreds of crAss-like phage genomes has highlighted the diversity and global prevalence of these phages in human gut microbiomes. CrAss-like phages have even been found in non-human primates[14], suggesting these phages have been evolving alongside humans for millions of years. However, it is currently unknown when and how an individual typically acquires crAss-like phages, as well as what level of strain diversity exists within the microbiome of an individual. The datasets examined here[15,16] contain mother-infant pairs sampled extensively during the first year of life and represent a unique opportunity to answer these questions.

We first characterized p-crAssphage and found no samples collected from infants within 24h of birth met our 1x coverage threshold. P-crAssphage becomes increasingly prevalent as infants age, but does not reach the levels found in mothers by one year of life. The host(s) of p-crAssphage may not be present or have reached sufficient abundance in some infants by the end of sample collection. Infants acquire many of their gut bacteria through direct transmission from their mother, while gut viromes have been shown to remain unique between family members and twins[38]. In contrast to other members of the gut virome, we found nearly identical assembled p-crAssphage genomes in 6/10 cases where mothers and infants both harbor the phage, suggesting vertical transmission. However, we cannot rule out alternative possibilities, such as transmission from a different family member or from a common environmental source. We also observed cases where mothers and infants had unrelated p-crAssphage genomes and cases where infants had p-crAssphage but it was undetectable in the mother, which argue that the infants acquired p-crAssphage from an undetectable population in the mother or from another source.

It is currently unknown if individuals are typically colonized by a single or multiple p-crAssphage strains, or how similar or different these strains may be. We characterized the strain

bioRxiv preprint doi: https://doi.org/10.1101/460113; this version posted October 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

diversity of the p-crAssphage population in an individual by examining positions in the p-crAssphage genome where we detected multiple high-quality alleles. We found most mothers have a limited number of variable sites, with a median frequency of 0.04% across the 97kb p-crAssphage genome, arguing that most mothers have a limited diversity of p-crAssphage strains. We did observe one mother with 100x more frequent variable sites, however. Infants generally have an order of magnitude fewer variable sites than mothers, suggesting a population that is further reduced in strain diversity, which may be the result of a bottleneck event upon acquisition or transmission. P-crAssphage is significantly less diverse than the second most abundant phage in these samples, *Lactococcus phages 16802*, where variable sites are detected with a median frequency of 5.9%. In cases where we observed putative mother-infant transmission, major alleles in the mother are primarily found in the infant, suggesting the mother's dominant strain is primarily responsible for colonizing the infant. The p-crAssphage population in infants develops additional variable sites over time, often at positions where only single alleles were detected in the mother. This could be due to the different bacterial hosts, nutritional sources and selective pressures in the infant microbiome, or simply due to random mutations.

In the most reductionist sense, two p-crAssphage genomes could differ at a single position and be considered different strains. However, we are most interested in strain variation that has functional consequences for the phage, its host or other members of the gut microbiome. Strain diversity in the p-crAssphage population of mothers is enriched for variants with predicted synonymous effects. However, we do observe enrichment for nonsynonymous (i.e. functional) variants in key genes, including predicted tail fiber proteins. This suggests that there may be a benefit to maintaining nonsynonymous allelic diversity in these genes, such as the ability to infect a broader range of hosts. One isolated crAss-like phage[20] was noted to have a very specific host range, so variation in tail fiber genes may allow these phage to infect an increased range of bacteria. Laboratory experiments are necessary to further investigate this

bioRxiv preprint doi: https://doi.org/10.1101/460113; this version posted October 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

hypothesis, but could use exiting variation in the tail genes as a starting point to screen for expanded host range. P-crAssphage in infants has variable sites that are enriched for synonymous changes compared to mothers, but limited sample numbers made it difficult to determine enrichment for specific genes.

P-crAssphage is the first described member of a diverse group of crAss-like phages[5], with four "family" level and ten "genus" level classifications. Similar to p-crAssphage, we observe a trend of increasing prevalence with infant age for many clusters of crAss-like phages. Some clusters, such as Alpha 03, are prevalent in mothers but rarely or never observed in infants, suggesting the hosts of these phages have yet to reach sufficient abundance in the infant microbiome. We first observe a crAss-like phage at 1x coverage in samples collected three days after birth. In the case of family 633, the mother and three-day old infant have a Delta 07 phage with 99.3% alignment identity. Since we did not observe such early potential transmission events with p-crAssphage, this may represent the first detectable transmission of any crAss-like phage from mother to infant, and a lower limit for the time for a crAss-like phage to colonize the infant microbiome. Alternatively, crAss-like phages may not colonize the infant microbiome at such an early time, rather, they may be acquired through routes other than actual parturition. For example, the phages might be present in yet understudied niches, such as the mother's breast milk or the shared built environment of the baby and mother. Of note, *Bacteroides species*, which are posited to be the natural host of p-crAssphage and are known to be the host of a crAss-like phage[20] have previously been detected in breast milk[39]. Overall, we find nearly identical genomes in 50% of cases when we detect the same cluster crAss-like phage in both mother and infant, suggesting a transmission rate similar to p-crAssphage.

Regardless of the crAssphage status of the mother, we found a strong association of p-crAssphage and crAss-like phage presence with vaginal delivery, in contrast to what has been described previously[7]. One potential explanation is that vaginal birth is responsible for transmitting the phage from mother to infant. However, this is less likely in cases where infants

bioRxiv preprint doi: https://doi.org/10.1101/460113; this version posted October 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

harbor a phage undetected in the mother. Another possible explanation is that vaginal birth is responsible for seeding bacteria necessary for later colonization by crAss-like phages. Previous research found maternal seeding of bacteria from the class Bacteroidia was inhibited by C-section birth, supporting this hypothesis[24,40]. Future research with more balanced cohorts will likely clarify whether or not birth mode affects crAss-like phage acquisition and transmission. Unexpectedly, microbiomes of vaginally-born infants positive for crAss-like phages were strongly enriched in *Collinsella species*. It is doubtful that this finding suggests new hosts for crAss-like phages, rather, Collinsella may be a hallmark for a developing and increasingly anaerobic infant microbiome that is capable of harboring these phages.

Finally, we observe that p-crAssphage is frequently transmitted via fecal microbiota transplantation (FMT) and can engraft stably in FMT recipients for up to one year. Engraftment of bacteria and phages has been well-studied in the case of FMT treatment for recurrent *Clostridium difficile* infection, and transplantation of p-crAssphage has been identified previously[35–37]. Our strain-level findings add new insight into the transmission of lytic bacteriophages. We assembled nearly identical genomes from both donors and recipients, highly suggestive of transmission of the specific p-crAssphage strain. Taken together, the results from both populations suggest that infants and patients receiving FMT have relatively unpopulated, naive microbiomes, providing an open niche for p-crAssphage to engraft into.

While this study suggests new principles about acquisition and transmission of crAss-like phages in the gut microbiome, it does have several limitations. First, we examined publicly available metagenomic data and were therefore limited to the available study cohort and sample size. In the mother-infant studies, stool samples were not collected from family members other than mothers, which could help determine other contributions to crAss-like phage acquisition in infants. Also, many infant birth samples were limited by low sequencing depth. As such, the estimates for acquisition and transmission presented here are likely underestimates. Sampling time points, processing techniques and study populations were different between the two

bioRxiv preprint doi: https://doi.org/10.1101/460113; this version posted October 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

studies, although both were conducted in Northern European individuals. Second, short-read sequencing data limited our ability to phase strain variants in the p-crAssphage genome. If the samples were resequenced with long-read sequencing approaches[41], we could obtain single reads spanning many variable sites. This would allow us to determine if the observed variants are the result of a smaller number of more divergent strain populations, or a high number of closely related strains. Finally, our group has become aware of false positive strain sharing results due to “barcode swapping” in dual-indexed Illumina sequencing libraries generated in our lab, which was first described in 2017[42]. As the indexing strategy was not reported for the public data we analyzed in this manuscript, we cannot be certain that the findings presented are not the result of this artifact. However, we believe our results, where only matched mother-infant pairs and matched FMT donor-recipient pairs share highly related crAss-like phage sequences, are unlikely to be explained by barcode swapping alone. The detrimental effect of barcode swapping also highlights the importance of reporting index sequences as a key part of making data publicly available.

Future work expanding on our findings should be directed towards answering several important questions. How stable are crAss-like phages transmitted from mother to infant over time? Are they lifelong inhabitants that, barring heavy antibiotic use, can be transmitted for generations? Are there exclusion principles that prevent the acquisition of a second, more divergent p-crAssphage strain? Additionally, our strain diversity analysis focused on p-crAssphage, but a wealth of diversity is also present in crAss-like phages. Better genome annotations and more concrete principles surrounding the identity and taxonomy of crAss-like phages will enable this research, and isolating, culturing and characterizing new crAss-like phages is a key next step. Finally, long-read metagenomic sequencing will enable better analysis of the strain populations among crAss-like phages in the mixed community of the microbiome. The ubiquity, distinct genome composition and ease of computational analysis with crAss-like phages may render them useful models for querying microbial transmission more

bioRxiv preprint doi: https://doi.org/10.1101/460113; this version posted October 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

broadly. Future work remains to determine precisely whether, and how, crAss-like phages influence the gut ecosystem and ultimately human health.

bioRxiv preprint doi: https://doi.org/10.1101/460113; this version posted October 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Methods

Sequence read preprocessing

Raw sequencing reads from Bäckhed et al.[15] and Yassour et al.[16] were downloaded from SRA from each sample and preprocessed in a consistent way: TrimGalore version 0.5.0[43] was used to perform quality and adapter trimming with the flags '--clip_R1 15 --clip_R2 15 --length 60'. SeqKit version 0.9.1[44] was used to remove duplicates with the command 'seqkit rmdup --by-seq'. Reads were mapped against the human genome using BWA version 0.7.17-r1188[45] and only unmapped reads were retained. Many infant samples at birth had low read counts after preprocessing, and samples with fewer than 10,000 reads were removed from all subsequent analyses. This left 135 families with sufficient depth in at least one sample from mother and infant.

Kraken2 classification

For classification of p-crAssphage, we built a Kraken2[21] database containing all bacteria, viral and fungal genomes in NCBI GenBank assembled to complete genome, chromosome or scaffold quality as of February 2019. Human and mouse reference genomes were also included in the database. A Bracken[46] database was also built with a read length of 150 and kmer length of 35. P-crAssphage is represented by a 97kb genome (accession NC_024711.1). Multiple crAss-like phages are present in GenBank and would cause reads mapping to multiple genomes to be classified at the least common ancestor of "crAss-like viruses." To prevent this from happening, other crAss-like genomes were removed from the database.

For classification of crAss-like phages, we added to the viral database, replacing the original "crAss-like viruses" clade with genomes in the proposed subfamily and cluster hierarchy

bioRxiv preprint doi: https://doi.org/10.1101/460113; this version posted October 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

described in Guerin et al.[5]. Kraken2 was used with default classification parameters on paired-end reads.

For testing associations between crAss-like phage presence and other bacterial taxa, reads were classified by using Kraken2 with default parameters on paired-end reads, and Bracken was used for abundance estimation with the parameters '-r 150 -l S -t 10'

Assembling and comparing crAss-like phage genomes

Preprocessed sequencing reads were assembled with SPAdes version 3.13.1[47] using the '--meta' flag. Contigs ≥500bp were aligned with BWA against either the p-crAsspahge reference genome or composite genomes from all the crAss-like phages in a cluster from Guerin et al.[5]. Resulting contigs were assessed for their N50 and total assembly length. Pairwise comparisons were conducted with nucmer version 4.0.0beta2[48] and the average identity and total length of 1-1 aligned segments was reported. The heatmap in Figure 1 was clustered on the euclidean distance between samples with the ward.d2 clustering method and plotted with the *heatmap.2* function in the *gplots* package for R[49].

SNPs and multiallelic sites

SNPs were called with Snippy[50] with freebayes[51] as the variant caller using the p-crAssphage reference at sites covered ≥10x. Filtering, decomposition and normalization of variants was necessary to compare between samples and was conducted with vt version 0.5[52] and bcftools version 1.9[53]. The output of snippy, *snps.raw.vcf,* was used in this command: "vt decompose -s snps.raw.vcf | vt decompose_blocksub -a - | bcftools norm -f crassphage_reference.fasta -m -any | bcftools view --include 'QUAL≥100 && FMT/DP≥10 && (FMT/AO)/(FMT/DP)≥0'". We

bioRxiv preprint doi: https://doi.org/10.1101/460113; this version posted October 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

calculated the transition/transversion ratio of detected variants using *vcftools*[54] version 0.1.16. Considering all detected variants agnostic of samples, the transition/transversion ratio is 2.91 for all SNPs, 2.41 for fixed SNPs after variant decomposition and 2.42 for multiallelic SNPs at >0.1 AF after variant decomposition. Considering samples individually, the median Ts/Tv ratios are 3.40, 2.80 and 3.23 (SD=2.3, 1.5, 3.4), respectively. The median sample numbers are higher because of samples with few detected transversions producing a comparatively high ratio.

To compare fixed SNPs between samples, we only consider sites covered ≥10x in both samples. The reported SNP % identity is 1 - (the number of fixed SNPs different between samples / number of sites covered ≥10x in both samples).

Multiallelic sites were called as sites with two alleles and ≥5 reads supporting each allele. We report a normalized statistic (F_{multi}) to compare multiallelic sites across samples with highly variable coverage. At a given minor allele fraction(AF), F_{multi} is the proportion of multiallelic sites with a minor AF > *x* among those sites covered well enough to detect a minor AF of *x.* Effects of multiallelic variants were predicted with SnpEff version 4.3[27] using the p-crAssphage genome annotation available on GenBank and the flags `ann -noLog -noStats -no-downstream -no-upstream -no-utr -t`. When mothers had multiple samples, we used the one with the highest p-crAssphage coverage for multiallelic site analysis.

CrAss-like phage correlation with bacterial abundance

We used the outputs of Bracken to test for differential abundances in taxa between groups. Matrices of reads classified to each taxa in each sample were filtered to keep only taxa with an abundance of at least 0.001 and nonzero values in at least 30% of samples. Zeros in the data were replaced with the Geometric Bayesian multiplicative method in the zCompositions version

bioRxiv preprint doi: https://doi.org/10.1101/460113; this version posted October 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

1.3.2-1 package for R[55]. Differential abundance between groups was calculated with the ALDEx2 version 1.16.0 package for R[33].

FMT data analysis

Data from the FMT studies[35,37] were processed, assembled and compared in the same way as mother/infant data. Sample MGH06R[35] was excluded from the FMT cohort analysis as it could not be definitively determined whether sample designated as pre-FMT was actually collected prior to transplantation (personal communication with authors).

Data availability

All data analyzed in this study are publicly available. We have deposited the p-crAssphage and crAss-like phage assembled genomes as a supplementary file along with this manuscript. Workflows for preprocessing, metagenomic assembly, assembly comparison and SNP calling in metagenomic data were used in this manuscript and can be found at:

https://github.com/bhattlab/bhattlab_workflows

A workflow for Kraken2 classification can be found at:

https://github.com/bhattlab/kraken2_classification

Acknowledgements

We thank Benjamin Good for advice on statistical tests related to figure 3, and Dylan Maghini for helpful comments on the manuscript. This work was supported in part by the National Science Foundation Graduate Research Fellowship and the Stanford CEHG Pre-Doctoral Fellowship (FBT) as well as the National Cancer Institute NIH K08 CA184420, the Damon Runyon Clinical

bioRxiv preprint doi: https://doi.org/10.1101/460113; this version posted October 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Investigator Award, and the Amy Strelzer Manasevit Award (ASB). This work was supported in part by the NIH grant P30 CA124435 which supports the following Stanford Cancer Institute Shared Resource: the Genetics Bioinformatics Service Center.

Author contributions

BS and FBT performed computational analysis. GS designed the multi-allelic site analysis and edited the manuscript. BS, FBT, and ASB designed the study and wrote and edited the manuscript.

Figure captions

Figure 1:

Mother-infant pairs share >99.7% similar p-crAssphage genomes in 6/10 cases. Heatmap of pairwise alignment % identity of metagenome-assembled p-crAssphage genomes from mothers and infants. Only families with p-crAssphage detected in at least one mother and infant sample are shown. The p-crAssphage reference genome is also included as a comparison.

Figure 2:

P-crAssphage populations in mothers and infants differ in multiallelic sites, and therefore strain diversity.

bioRxiv preprint doi: https://doi.org/10.1101/460113; this version posted October 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

a. The p-crAssphage population in mothers has more multiallelic sites than the p-crAssphage population in infants. F_{multi} (fraction of the p-crAssphage genome with multiallelic sites detected at the given allelic fraction threshold) in all mother and infant samples with at least one multiallelic site detected. P-values were calculated with the two-sided Wilcoxon rank sum test and are uncorrected for multiple hypothesis testing.

P-crAssphage populations in mothers do not change in the number of multiallelic sites over time. F_{multi} for mother samples from Yassour et al.[16]. P-values were calculated with a linear mixed model to account for repeated sampling of the same individual.

Allelic fraction of multiallelic sites in the p-crAssphage genome from mothers that are fixed in her infant. The distribution is separated by alleles that are present in the infant's p-crAssphage or not. P-values were calculated with the two-sided Wilcoxon rank sum test.

Schematic depicting multiallelic sites in mother and infant samples over time. In the three cases where p-crAssphage was detected in the mother and multiple samples from the same infant, infants develop more multiallelic sites over time.

Figure 3:

The p-crAssphage population in mothers has limited strain diversity, and predicted effects of multiallelic sites differ in the p-crAssphage population of mothers and infants.

a. Distribution of multiallelic sites per kilobase in samples from mothers.

Distribution of predicted effects of multiallelic sites from mother and infant samples, compared to a background distribution of equal probability of each DNA change at each position in the p-crAssphage reference genome.

Figure 4:

bioRxiv preprint doi: https://doi.org/10.1101/460113; this version posted October 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The frequency and predicted effects of multiallelic sites vary across the p-crAssphage genome, and hotspots correspond to certain predicted genes. Track view of multiallelic sites (AF > 0.1) in samples from mothers, calculated with a sliding window of size 1500bp with step size 200.

a. Fraction of samples covered at least 10x.

%GC content of the p-crAssphage reference genome.

GC skew of the p-crAssphage reference genome.

Total count of multiallelic sites in the window.

Log base 2 ratio of nonsynonymous sites (N) to synonymous sites (S).

Annotation and selected predicted functions of genes in the reference genome.

Figure 5:

P-crAssphage status in patients receiving FMT over time. All patients were p-crAsphage negative prior to FMT, and several patients acquired p-crAssphage during the course of treatment.

a. P-crAssphage detection at 1x coverage in samples from Smillie et al.[35] Both donors shown were p-crAssphage positive.

P-crAssphage detection at 10x coverage in samples from Draper et al.[37] Donor D1 was p-crAssphage positive, while donors D2 and D3 were p-crAssphage negative

Figure S1:

P-crAssphage presence at 1x coverage in infant samples (a) increases over time, but does not reach the levels of mothers (b) by the end of sampling.

Figure S2:

P-crAssphage is more related in samples from mother-infant pairs than in samples from unrelated individuals.

bioRxiv preprint doi: https://doi.org/10.1101/460113; this version posted October 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

a. Distribution of pairwise alignment % identity of metagenome-assembled p-crAssphage genomes. Groups are separated by family relationships. P-values were calculated with the two-sided Wilcoxon rank sum test.

Distribution of pairwise SNP % identity of p-crAssphage genomes. Groups are separated by family relationships. P-values were calculated with the two-sided Wilcoxon rank sum test.

Figure S3:

Metagenome-assembled p-crAssphage genomes are highly similar in samples from matched mother-infant pairs and samples from the same individual. The heatmap shows pairwise alignment % identity in all mother-infant samples with p-crAssphage detected. The p-crAssphage reference genome is also included as a comparison.

Figure S4:

P-crAssphage is highly similar at the SNP level in samples from matched mother-infant pairs and samples from the same individual. The heatmap shows pairwise SNP % identity in all samples with p-crAssphage detected are shown.

Figure S5:

CrAss-like phages are detected in varying levels in mother and infant samples. Crass-like phage clusters are as defined in Guerin et al.[5]

a. CrAss-like phages detected at 1x coverage in samples from mothers in each study

b. CrAss-like phages detected at 1x coverage in samples from infants in each study.

c. Number of crAss-like phage clusters detected in each sample from mothers and infants

Figure S6:

bioRxiv preprint doi: https://doi.org/10.1101/460113; this version posted October 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Collinsella and *Collinsella aerofaciens* are at higher relative abundances in crAss-like phage positive vaginally delivered infants at 3-4 months of age, compared to crAss-like phage negative infants. P-values calculated with the two-sided Wilcoxon rank sum test and corrected for multiple hypothesis testing.

Figure S7:

Metagenome-assembled p-crAssphage genomes are highly similar in samples from matched FMT donor-recipient pairs in Smillie et al.[35]. The heatmap shows pairwise alignment % identity in all samples that assembled >50kb p-crAssphage sequence. The p-crAssphage reference genome is also included as a comparison.

Figure S8:

Metagenome-assembled p-crAssphage genomes are highly similar in samples from matched FMT donor-recipient pairs in Draper et al.[37]. The heatmap shows pairwise alignment % identity in all samples that assembled >50kb p-crAssphage sequence. The p-crAssphage reference genome is also included as a comparison. The donor for patient P7 was p-crAssphage negative; this patient may have acquired their p-crAssphage from the environment or another source.

Figure S9:

Lactococcus phages were the only other group of phages detected with at least 1x coverage in at least ten mother and infant samples. This best represented member of this group, *Lactococcus phage 16802,* was detected in 34 samples and has more multiallelic sites than p-crAssphage on average, with a median of 58.8 multiallelic sites per kb.

bioRxiv preprint doi: https://doi.org/10.1101/460113; this version posted October 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table captions

Table S1:

Sheet 1) Sample and dataset metadata for mother-infant samples.

Sheet 2) Sample and dataset metadata for FMT samples.

Sheet 3) Associations between birth mode and crAss-like phage detection. Only comparisons with at least ten crAss-like phage positive infants across the two studies are shown. P-values were calculated with Fisher's exact test and are uncorrected for multiple hypothesis testing.

Table S2:

Sheet 1) Summary statistics of p-crAssphage metagenomic assembly in mother and infant samples.

Sheet 2) Individual sample statistics of p-crAssphage metagenomic assembly in mother and infant samples.

Table S3:

Sheet 1) Pairwise alignment comparisons between metagenomic-assembled p-crAssphage genomes in mother and infant samples.

Table S4:

Sheet 1) Multiallelic sites detected in p-crAssphage genes in samples from mothers.

Sheet 2) Multiallelic sites detected in p-crAssphage genes in samples from infants.

Table S5:

Sheet 1) Presence of crAss-like phage clusters in samples from mothers and infants.

bioRxiv preprint doi: https://doi.org/10.1101/460113; this version posted October 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Sheet 2) Summary statistics of crass-like phage metagenomic assembly in mother and infant samples.

Sheet 3) Individual sample statistics of crAss-like phage metagenomic assembly in mother and infant samples.

Sheet 4) Pairwise alignment comparisons between metagenomic-assembled crAss-like phage genomes in mother and infant samples.

Table S6:

Sheet 1) Table header explanation for sheet 2.

Sheet 2) Associations between crAss-like phage presence and bacterial relative abundances in vaginally born infants of 3-4 months of age.

References

1. Touchon, M., Moura de Sousa, J. A. & Rocha, E. P. Embracing the enemy: the diversification of microbial gene repertoires by phage-mediated horizontal gene transfer. *Curr. Opin. Microbiol.* **38**, 66–73 (2017).
2. Cadwell, K. *et al.* Virus-plus-susceptibility gene interaction determines Crohn's disease gene Atg16L1 phenotypes in intestine. *Cell* **141**, 1135–1145 (2010).
3. Zhao, G. *et al.* Intestinal virome changes precede autoimmunity in type I diabetes-susceptible children. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E6166–E6175 (2017).
4. Dutilh, B. E. *et al.* A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nat. Commun.* **5**, 4498 (2014).
5. Guerin, E. *et al.* Biology and Taxonomy of crAss-like Bacteriophages, the Most Abundant Virus in the Human Gut. *Cell Host Microbe* (2018). doi:10.1016/j.chom.2018.10.002

bioRxiv preprint doi: https://doi.org/10.1101/460113; this version posted October 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

6. Honap, T. P. *et al.* Biogeographic study of human gut associated crAssphage suggests impacts from industrialization and recent expansion. *bioRxiv* 384677 (2018). doi:10.1101/384677

7. McCann, A. *et al.* Viromes of one year old infants reveal the impact of birth mode on microbiome diversity. *PeerJ* **6**, e4694 (2018).

8. Cervantes-Echeverría, M. *et al.* Whole-genome of Mexican-crAssphage isolated from the human gut microbiome. *BMC Res. Notes* **11**, 902 (2018).

9. Stachler, E., Akyon, B., de Carvalho, N. A., Ference, C. & Bibby, K. Correlation of crAssphage qPCR Markers with Culturable and Molecular Indicators of Human Fecal Pollution in an Impacted Urban Watershed. *Environ. Sci. Technol.* **52**, 7505–7512 (2018).

10. Liang, Y., Jin, X., Huang, Y. & Chen, S. Development and application of a real-time polymerase chain reaction assay for detection of a novel gut bacteriophage (crAssphage). *J. Med. Virol.* **90**, 464–468 (2018).

11. Cinek, O. *et al.* Quantitative CrAssphage real-time PCR assay derived from data of multiple geographically distant populations. *J. Med. Virol.* **90**, 767–771 (2018).

12. Yutin, N. *et al.* Discovery of an expansive bacteriophage family that includes the most abundant viruses from the human gut. *Nat Microbiol* **3**, 38–46 (2018).

13. Liang, Y. Y., Zhang, W., Tong, Y. G. & Chen, S. P. crAssphage is not associated with diarrhoea and has high genetic diversity. *Epidemiol. Infect.* **144**, 3549–3553 (2016).

14. Edwards, R. A. *et al.* Global phylogeography and ancient evolution of the widespread human gut virus crAssphage. *Nature Microbiology* 1 (2019).

15. Bäckhed, F. *et al.* Dynamics and Stabilization of the Human Gut Microbiome during the First Year of Life. *Cell Host Microbe* **17**, 852 (2015).

16. Yassour, M. *et al.* Strain-Level Analysis of Mother-to-Child Bacterial Transmission during the First Few Months of Life. *Cell Host Microbe* **24**, 146–154.e4 (2018).

bioRxiv preprint doi: https://doi.org/10.1101/460113; this version posted October 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

17. Ferretti, P. *et al.* Mother-to-Infant Microbial Transmission from Different Body Sites Shapes the Developing Infant Gut Microbiome. *Cell Host Microbe* **24**, 133–145.e5 (2018).
18. Dominguez-Bello, M. G. *et al.* Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 11971–11975 (2010).
19. Reyes, A. *et al.* Viruses in the faecal microbiota of monozygotic twins and their mothers. *Nature* **466**, 334–338 (2010).
20. Shkoporov, A. N. *et al.* ΦCrAss001 represents the most abundant bacteriophage family in the human gut and infects Bacteroides intestinalis. *Nat. Commun.* **9**, 4781 (2018).
21. Wood, D. E. & Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **15**, R46 (2014).
22. Lim, E. S. *et al.* Early life dynamics of the human gut virome and bacterial microbiome in infants. *Nat. Med.* **21**, 1228–1234 (2015).
23. Jakobsson, H. E. *et al.* Decreased gut microbiota diversity, delayed Bacteroidetes colonisation and reduced Th1 responses in infants delivered by caesarean section. *Gut* **63**, 559–566 (2014).
24. Shao, Y. *et al.* Stunted microbiota and opportunistic pathogen colonization in caesarean-section birth. *Nature* **574**, 117–121 (2019).
25. Liu, Y. *et al.* The Perturbation of Infant Gut Microbiota Caused by Cesarean Delivery Is Partially Restored by Exclusive Breastfeeding. *Front. Microbiol.* **10**, 598 (2019).
26. Asnicar, F. *et al.* Studying Vertical Microbiome Transmission from Mothers to Infants by Strain-Level Metagenomic Profiling. *mSystems* **2**, (2017).
27. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly* **6**, 80–92 (2012).
28. Goz, E., Mioduser, O., Diament, A. & Tuller, T. Evidence of translation efficiency adaptation of the coding regions of the bacteriophage lambda. *DNA Res.* **24**, 333–342 (2017).

bioRxiv preprint doi: https://doi.org/10.1101/460113; this version posted October 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

29. Ghequire, M. G. K. *et al.* Different Ancestries of R Tailocins in Rhizospheric Pseudomonas Isolates. *Genome Biol. Evol.* **7**, 2810–2828 (2015).

30. Benler, S. *et al.* A diversity-generating retroelement encoded by a globally ubiquitous Bacteroides phage. *Microbiome* **6**, 191 (2018).

31. Scholl, D., Rogers, S., Adhya, S. & Merril, C. R. Bacteriophage K1-5 encodes two different tail fiber proteins, allowing it to infect and replicate on both K1 and K5 strains of Escherichia coli. *J. Virol.* **75**, 2509–2515 (2001).

32. Brown, B. P. *et al.* crAssphage abundance and genomic selective pressure correlate with altered bacterial abundance in the fecal microbiota of South African mother-infant dyads. doi:10.1101/582015

33. Fernandes, A. D. *et al.* Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. *Microbiome* **2**, 15 (2014).

34. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)* **57**, 289–300 (1995).

35. Smillie, C. S. *et al.* Strain Tracking Reveals the Determinants of Bacterial Engraftment in the Human Gut Following Fecal Microbiota Transplantation. *Cell Host Microbe* **23**, 229–240.e5 (2018).

36. Moss, E. L. *et al.* Long-term taxonomic and functional divergence from donor bacterial strains following fecal microbiota transplantation in immunocompromised patients. *PLoS One* **12**, e0182585 (2017).

37. Draper, L. A. *et al.* Long-term colonisation with donor bacteriophages following successful faecal microbial transplantation. *Microbiome* **6**, 220 (2018).

38. Reyes, A. *et al.* Viruses in the faecal microbiota of monozygotic twins and their mothers. *Nature* **466**, 334–338 (2010).

bioRxiv preprint doi: https://doi.org/10.1101/460113; this version posted October 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

39. Jiménez, E. *et al.* Metagenomic Analysis of Milk of Healthy and Mastitis-Suffering Women. *J. Hum. Lact.* **31**, 406–415 (2015).

40. Korpela, K. *et al.* Selective maternal seeding and environment shape the human gut microbiome. *Genome Res.* **28**, 561–568 (2018).

41. Moss, E. L. & Bhatt, A. S. Generating closed bacterial genomes from long-read nanopore sequencing of microbiomes. doi:10.1101/489641

42. Minimize index hopping in multiplexed runs. *Illumina* Available at: https://www.illumina.com/science/education/minimizing-index-hopping.html.

43. Krueger, F. Trim Galore! Available at: http://www.bioinformatics.babraham.ac.uk/projects/trim_galore.

44. Shen, W., Le, S., Li, Y. & Hu, F. SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. *PLoS One* **11**, e0163962 (2016).

45. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

46. Lu, J., Breitwieser, F. P., Thielen, P. & Salzberg, S. L. Bracken: estimating species abundance in metagenomics data. *PeerJ Computer Science* **3**, e104 (2017).

47. Nurk, S. *et al.* Assembling single-cell genomes and mini-metagenomes from chimeric MDA products. *J. Comput. Biol.* **20**, 714–737 (2013).

48. Marçais, G. *et al.* MUMmer4: A fast and versatile genome alignment system. *PLoS Comput. Biol.* **14**, e1005944 (2018).

49. Warnes, G. R. *et al.* gplots: Various R Programming Tools for Plotting Data.

50. Seeman, T. *snippy: fast bacterial variant calling from NGS reads*. (2015).

51. Garrison, E. & G., M. Haplotype-based variant detection from short-read sequencing. *arXiv preprint* (2012).

52. Tan, A., Abecasis, G. R. & Kang, H. M. Unified representation of genetic variants. *Bioinformatics* **31**, 2202–2204 (2015).

bioRxiv preprint doi: https://doi.org/10.1101/460113; this version posted October 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

53. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

54. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).

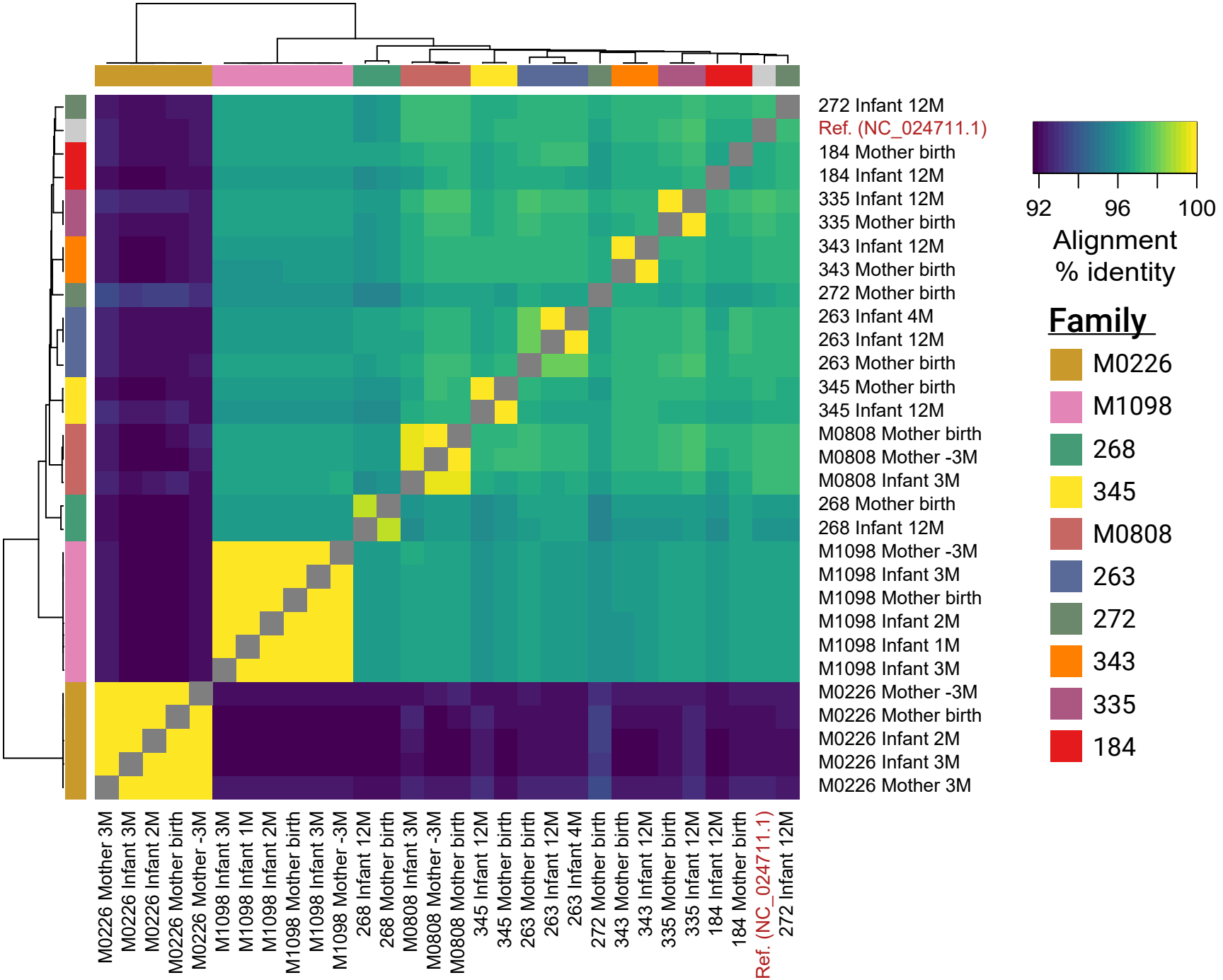
55. Palarea-Albaladejo, J. & Martín-Fernández, J. A. zCompositions — R package for multivariate imputation of left-censored data under a compositional approach. *Chemometrics and Intelligent Laboratory Systems* **143**, 85–96 (2015).

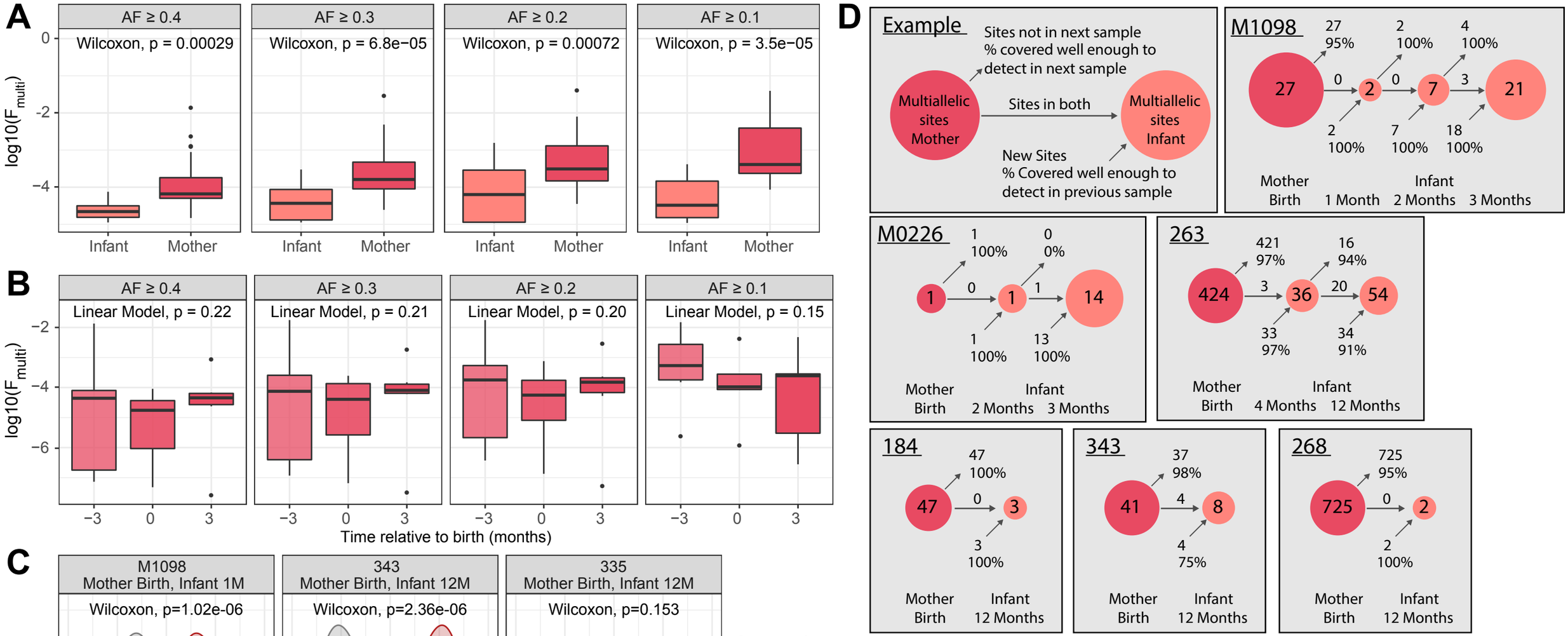
56. Yassour, M. *et al.* Natural history of the infant gut microbiome and impact of antibiotic treatment on bacterial strain diversity and stability. *Sci. Transl. Med.* **8**, 343ra81 (2016).

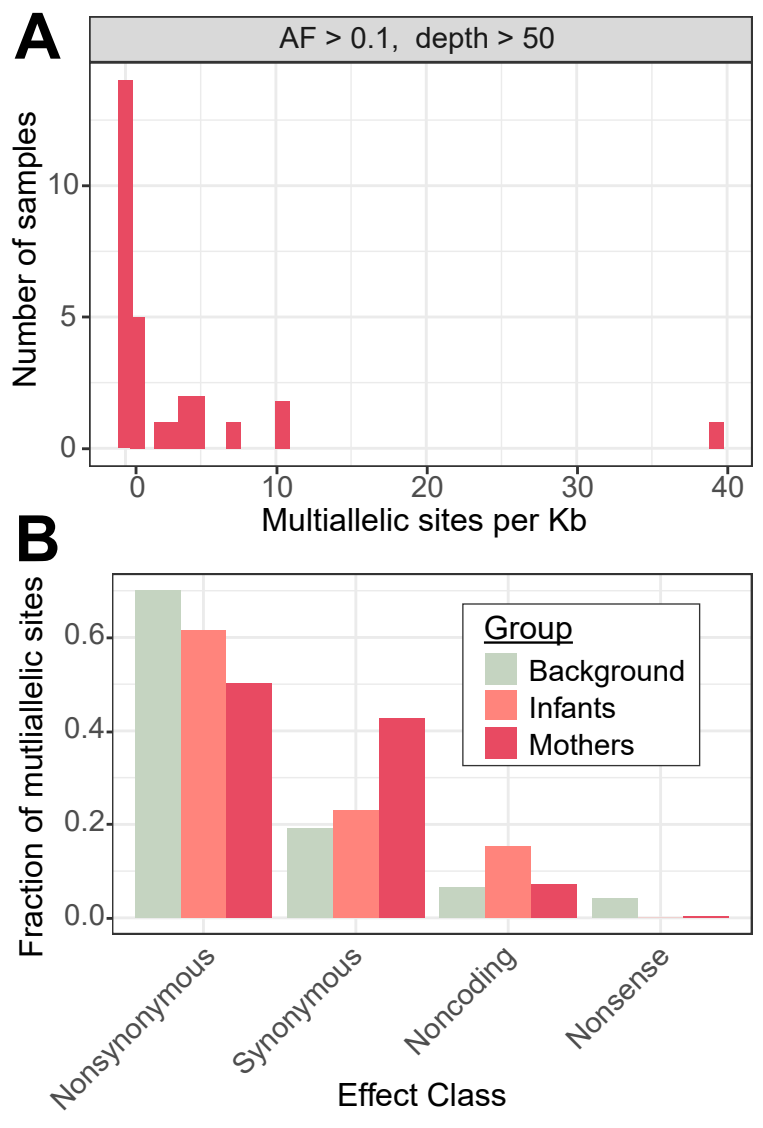
57. Chu, D. M. *et al.* Maturation of the infant microbiome community structure and function across multiple body sites and in relation to mode of delivery. *Nat. Med.* **23**, 314–326 (2017).

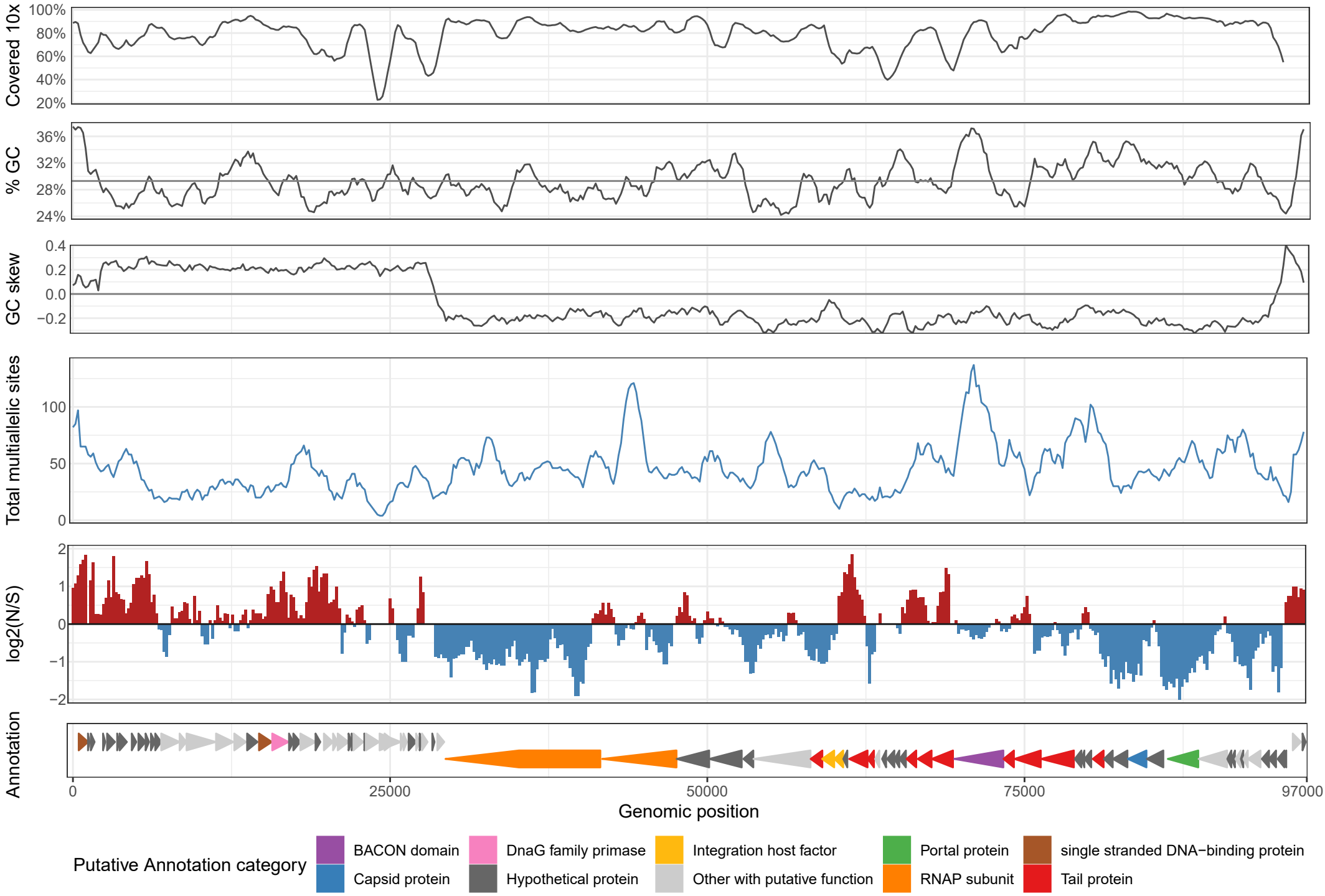
58. Minot, S. *et al.* Rapid evolution of the human gut virome. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 12450–12455 (2013).

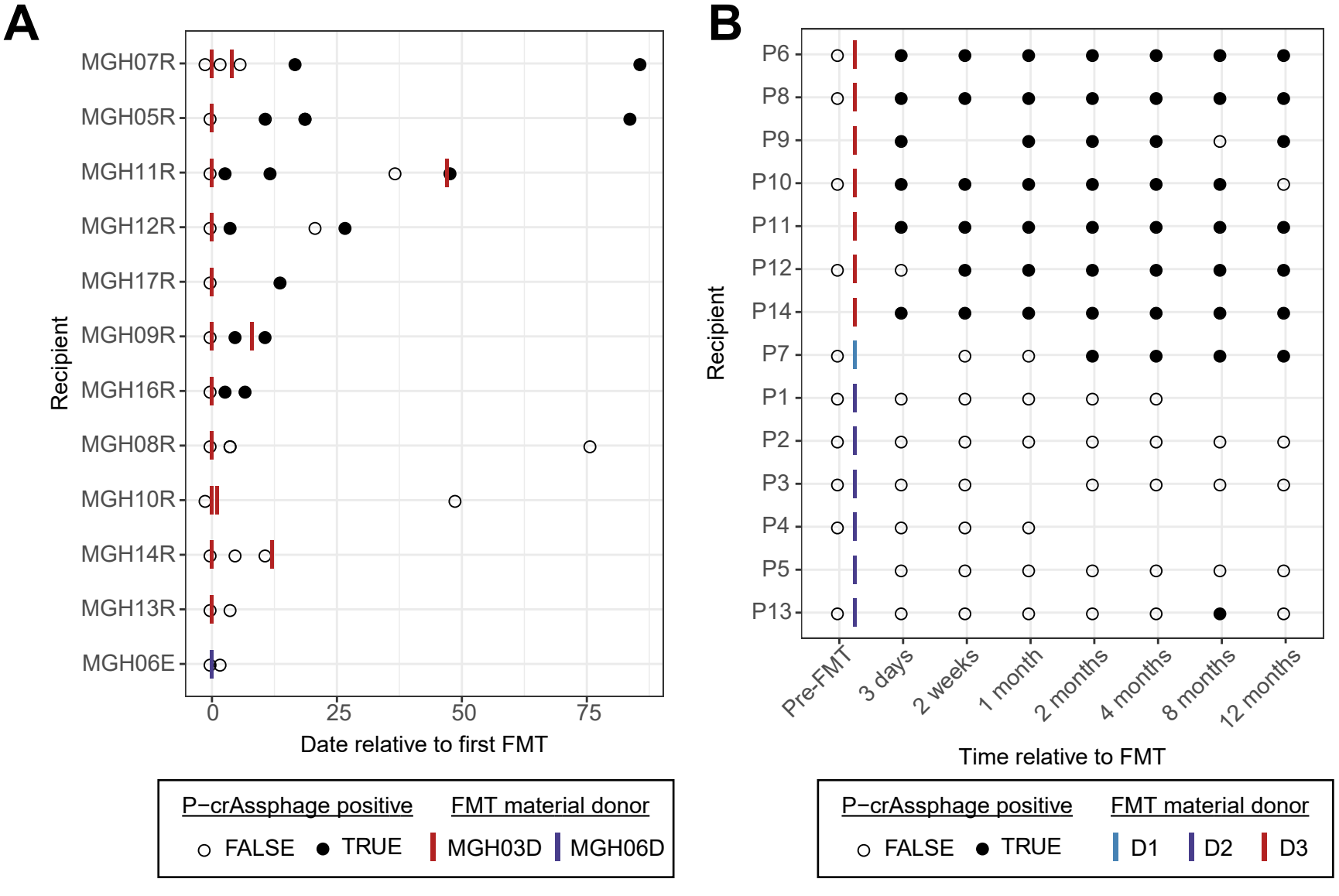
59. Minot, S. *et al.* The human gut virome: Inter-individual variation and dynamic response to diet. *Genome Res.* **21**, 1616–1625 (2011).

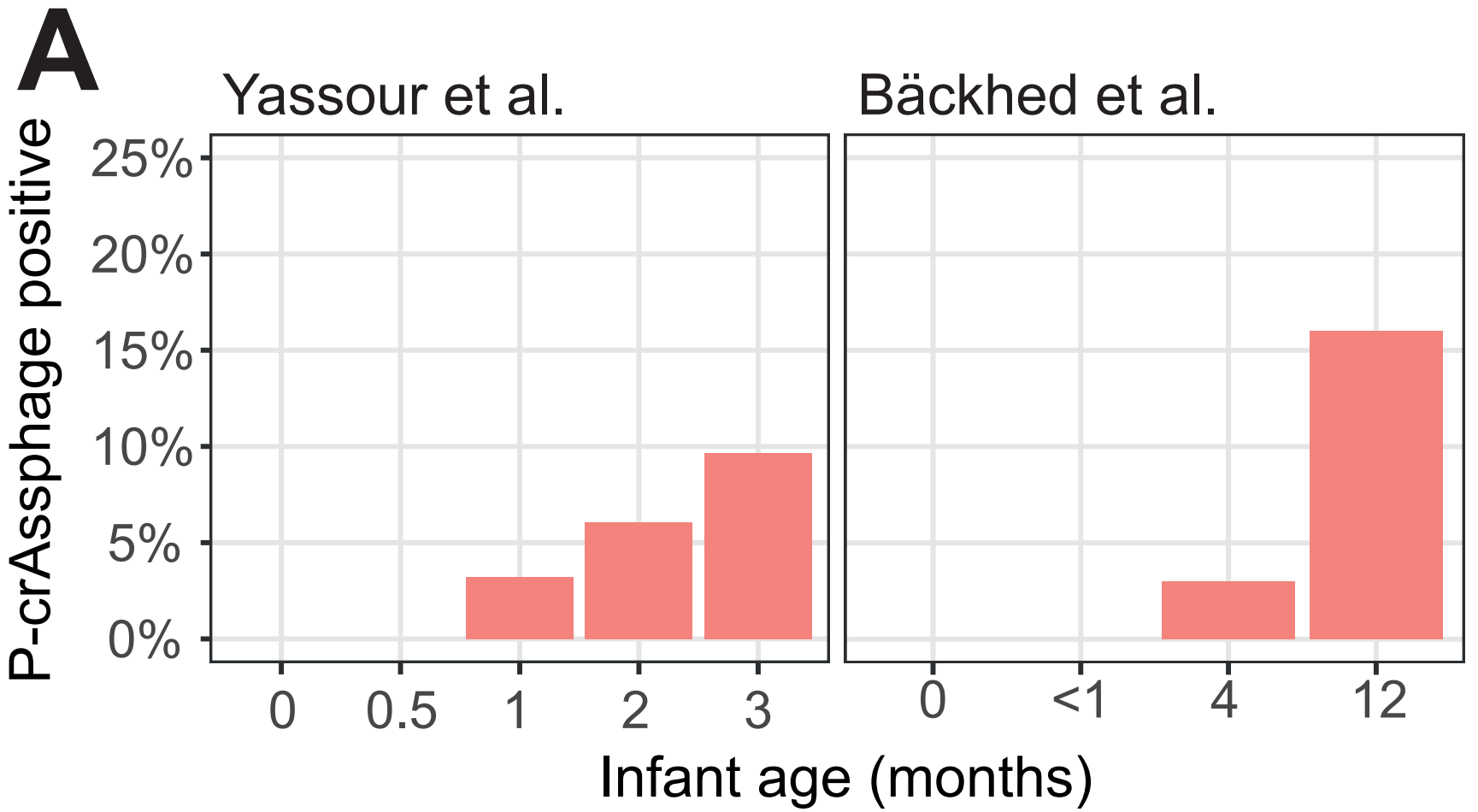

272 Infant 12M
Ref. (NC_024711.1)
184 Mother birth
184 Infant 12M
335 Infant 12M
335 Mother birth
343 Infant 12M
343 Mother birth
272 Mother birth
263 Infant 4M
263 Infant 12M
263 Mother birth
345 Mother birth
345 Infant 12M
M0808 Mother birth
M0808 Mother -3M
M0808 Infant 3M
268 Mother birth
268 Infant 12M
M1098 Mother -3M
M1098 Infant 3M
M1098 Mother birth
M1098 Infant 2M
M1098 Infant 1M
M1098 Infant 3M
M0226 Mother -3M
M0226 Mother birth
M0226 Infant 2M
M0226 Infant 3M
M0226 Mother 3M
M0226 Mother 3M
M0226 Infant 3M
M0226 Infant 2M
M0226 Mother birth
M0226 Mother -3M
M1098 Infant 3M
M1098 Infant 1M
M1098 Infant 2M
M1098 Mother birth
M1098 Infant 3M
M1098 Mother -3M
268 Infant 12M
268 Mother birth
M0808 Infant 3M
M0808 Mother -3M
M0808 Mother birth
345 Infant 12M
345 Mother birth
263 Mother birth
263 Infant 12M
263 Infant 4M
272 Mother birth
343 Mother birth
343 Infant 12M
335 Mother birth
335 Infant 12M
184 Infant 12M
184 Mother birth
Ref. (NC_024711.1)
272 Infant 12M
92
96
100
Alignment
% identity
Family
M0226
M1098
268
345
M0808
263
272
343
335
184

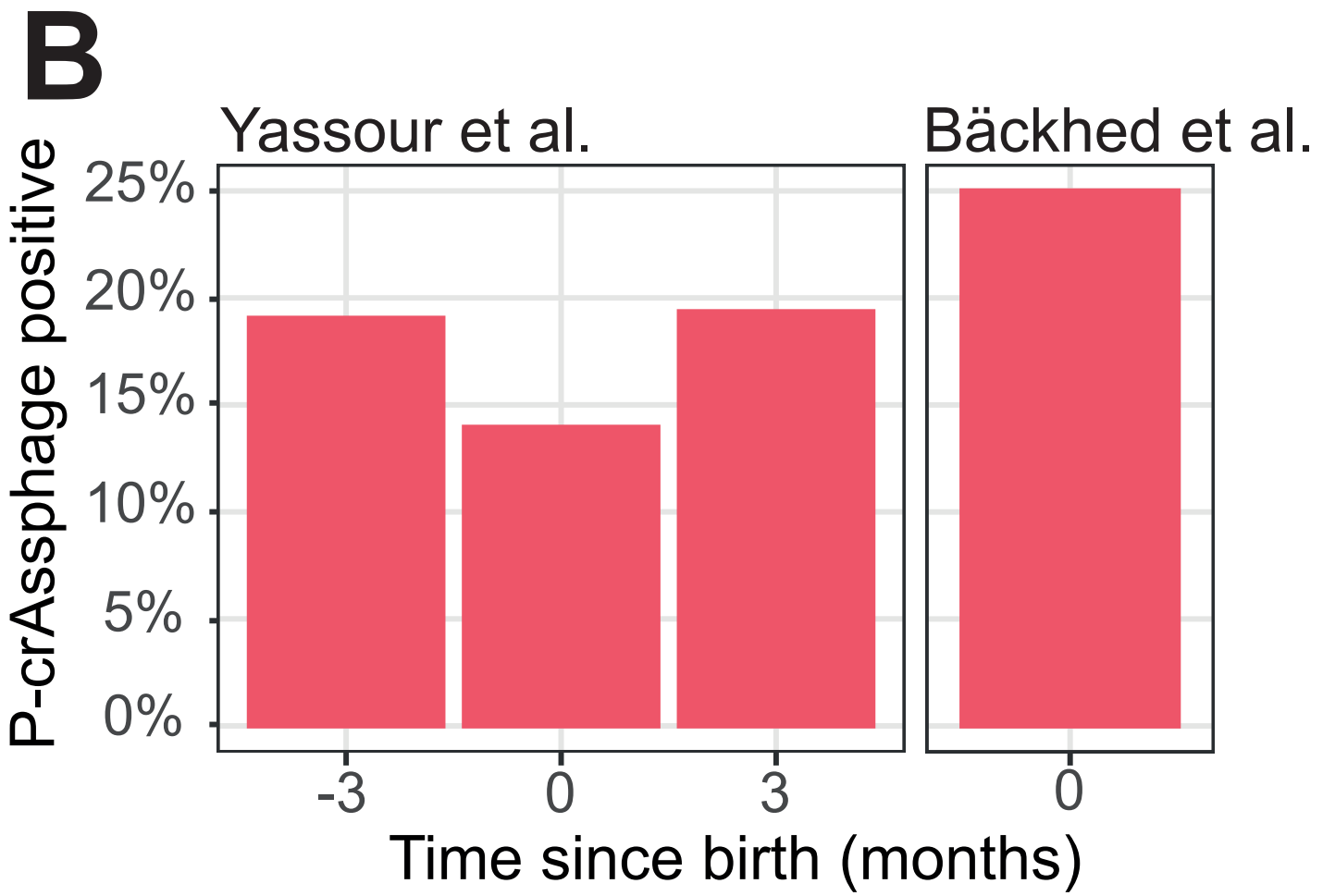

A
AF ≥ 0.4
Wilcoxon, p = 0.00029
AF ≥ 0.3
Wilcoxon, p = 6.8e−05
AF ≥ 0.2
Wilcoxon, p = 0.00072
AF ≥ 0.1
Wilcoxon, p = 3.5e−05
log10(F_multi)
Infant
Mother
B
AF ≥ 0.4
Linear Model, p = 0.22
AF ≥ 0.3
Linear Model, p = 0.21
AF ≥ 0.2
Linear Model, p = 0.20
AF ≥ 0.1
Linear Model, p = 0.15
log10(F_multi)
Time relative to birth (months)
C
M1098
Mother Birth, Infant 1M
Wilcoxon, p=1.02e-06
343
Mother Birth, Infant 12M
Wilcoxon, p=2.36e-06
335
Mother Birth, Infant 12M
Wilcoxon, p=0.153
Density
Allelic fraction in Mother
Site in infant
True
False
D
Example
Sites not in next sample
% covered well enough to detect in next sample
Multiallelic sites Mother
Sites in both
Multiallelic sites Infant
New Sites
% Covered well enough to detect in previous sample
M1098
27
95%
2
100%
4
100%
27
0
2
0
7
3
21
2
100%
7
100%
18
100%
Mother Birth
Infant
1 Month
2 Months
3 Months
M0226
1
100%
0
0%
1
0
1
1
14
1
100%
13
100%
Mother Birth
Infant
2 Months
3 Months
263
421
97%
16
94%
424
3
36
20
54
33
97%
34
91%
Mother Birth
Infant
4 Months
12 Months
184
47
100%
47
0
3
3
100%
Mother Birth
Infant
12 Months
343
37
98%
41
4
8
4
75%
Mother Birth
Infant
12 Months
268
725
95%
725
0
2
2
100%
Mother Birth
Infant
12 Months

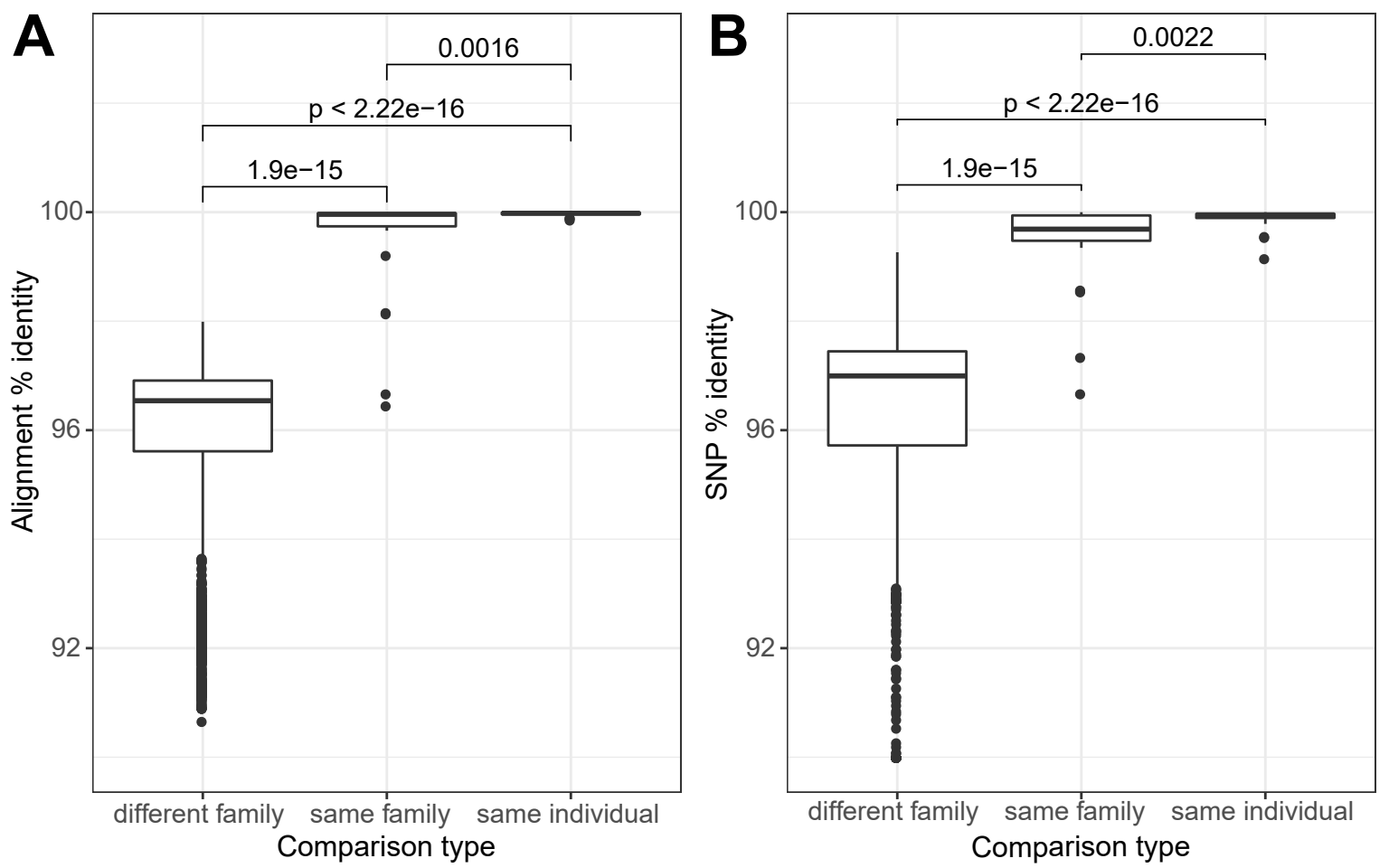

A
AF > 0.1, depth > 50
Number of samples
0
5
10
0
10
20
30
40
Multiallelic sites per Kb
B
Fraction of mutliallelic sites
0.0
0.2
0.4
0.6
Group
Background
Infants
Mothers
Nonsynonymous
Synonymous
Noncoding
Nonsense
Effect Class

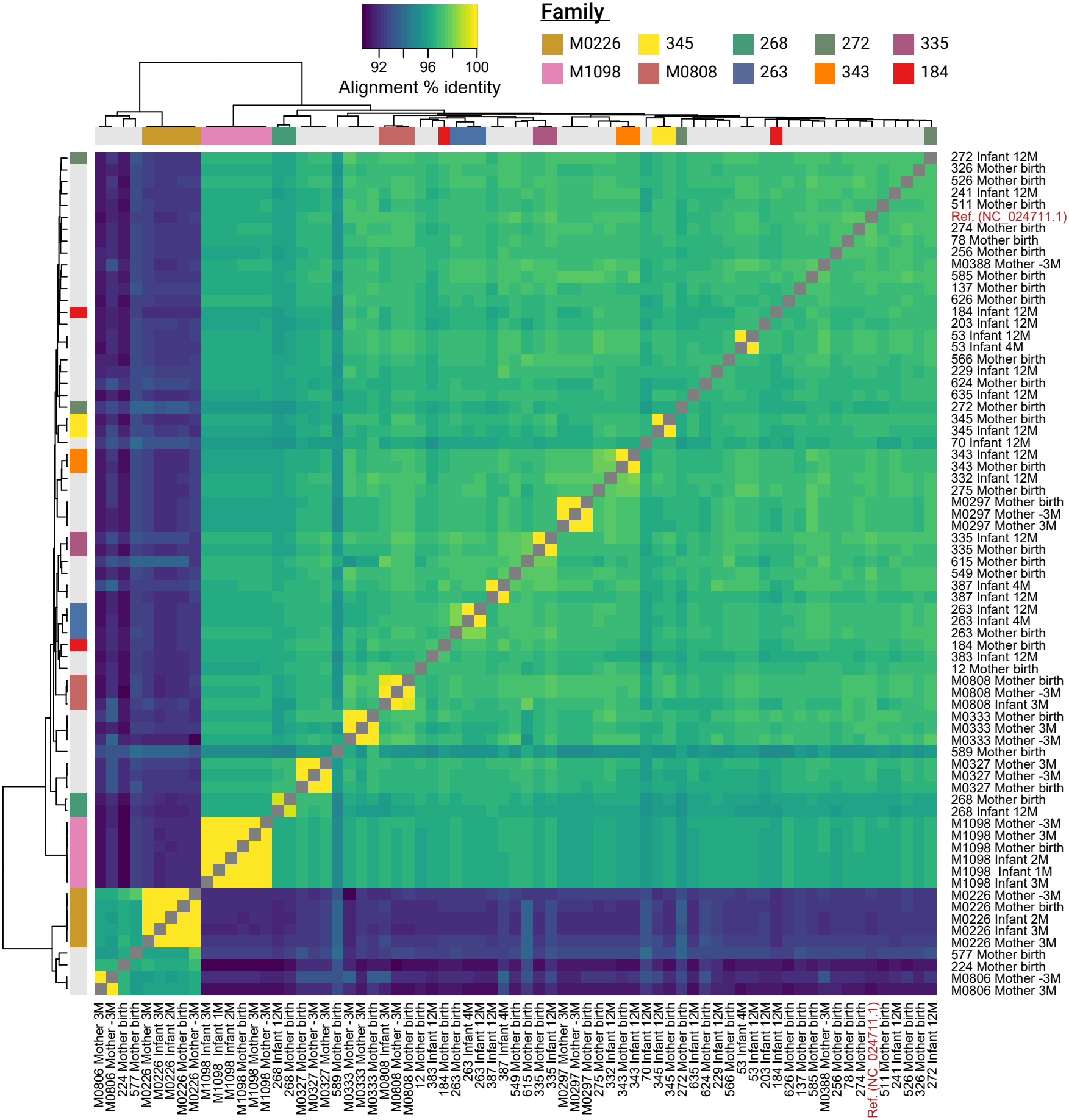

Covered 10x
100%
80%
60%
40%
20%
% GC
36%
32%
28%
24%
GC skew
0.4
0.2
0.0
−0.2
Total multiallelic sites
100
50
0
log2(N/S)
2
1
0
−1
−2
Annotation
0
25000
50000
75000
97000
Genomic position
Putative Annotation category
BACON domain
Capsid protein
DnaG family primase
Hypothetical protein
Integration host factor
Other with putative function
Portal protein
RNAP subunit
single stranded DNA−binding protein
Tail protein

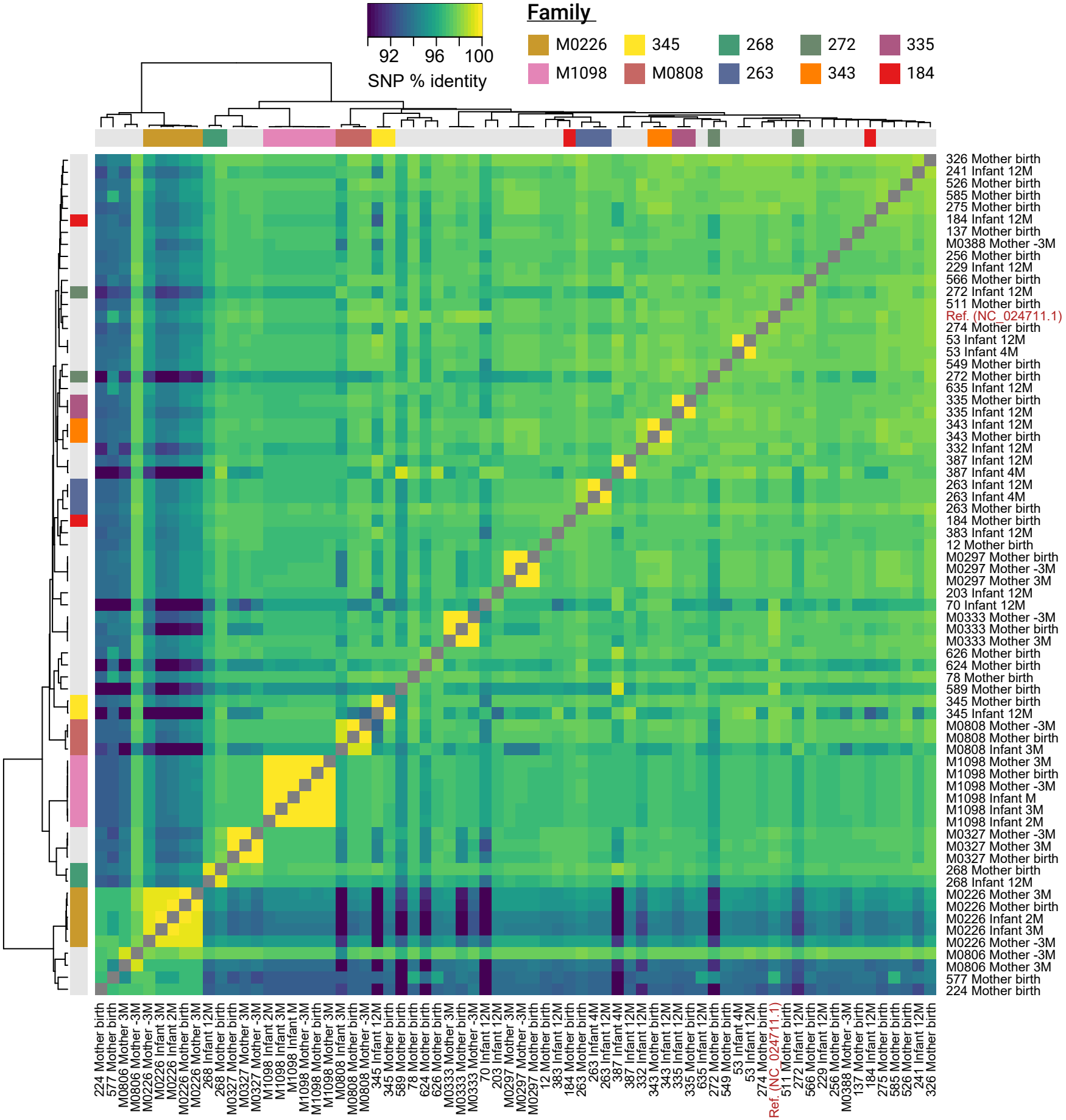

A
MGH07R
MGH05R
MGH11R
MGH12R
MGH17R
MGH09R
MGH16R
MGH08R
MGH10R
MGH14R
MGH13R
MGH06E
Recipient
0
25
50
75
Date relative to first FMT
P−crAssphage positive
FALSE
TRUE
FMT material donor
MGH03D
MGH06D
B
P6
P8
P9
P10
P11
P12
P14
P7
P1
P2
P3
P4
P5
P13
Recipient
Pre-FMT
3 days
2 weeks
1 month
2 months
4 months
8 months
12 months
Time relative to FMT
P−crAssphage positive
FALSE
TRUE
FMT material donor
D1
D2
D3

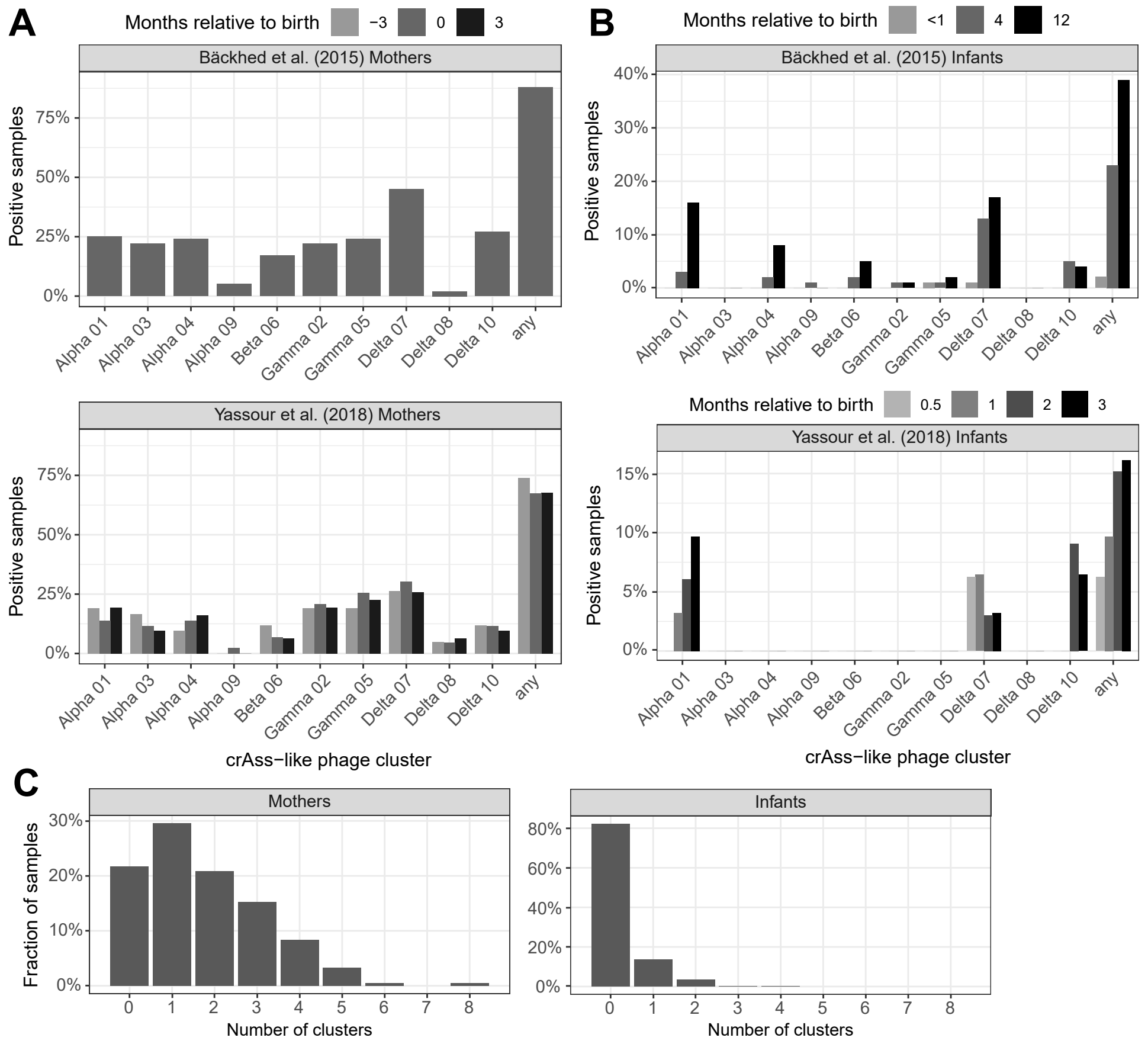

A
Yassour et al.
Bäckhed et al.
P-crAssphage positive
25%
20%
15%
10%
5%
0%
0
0.5
1
2
3
0
<1
4
12
Infant age (months)

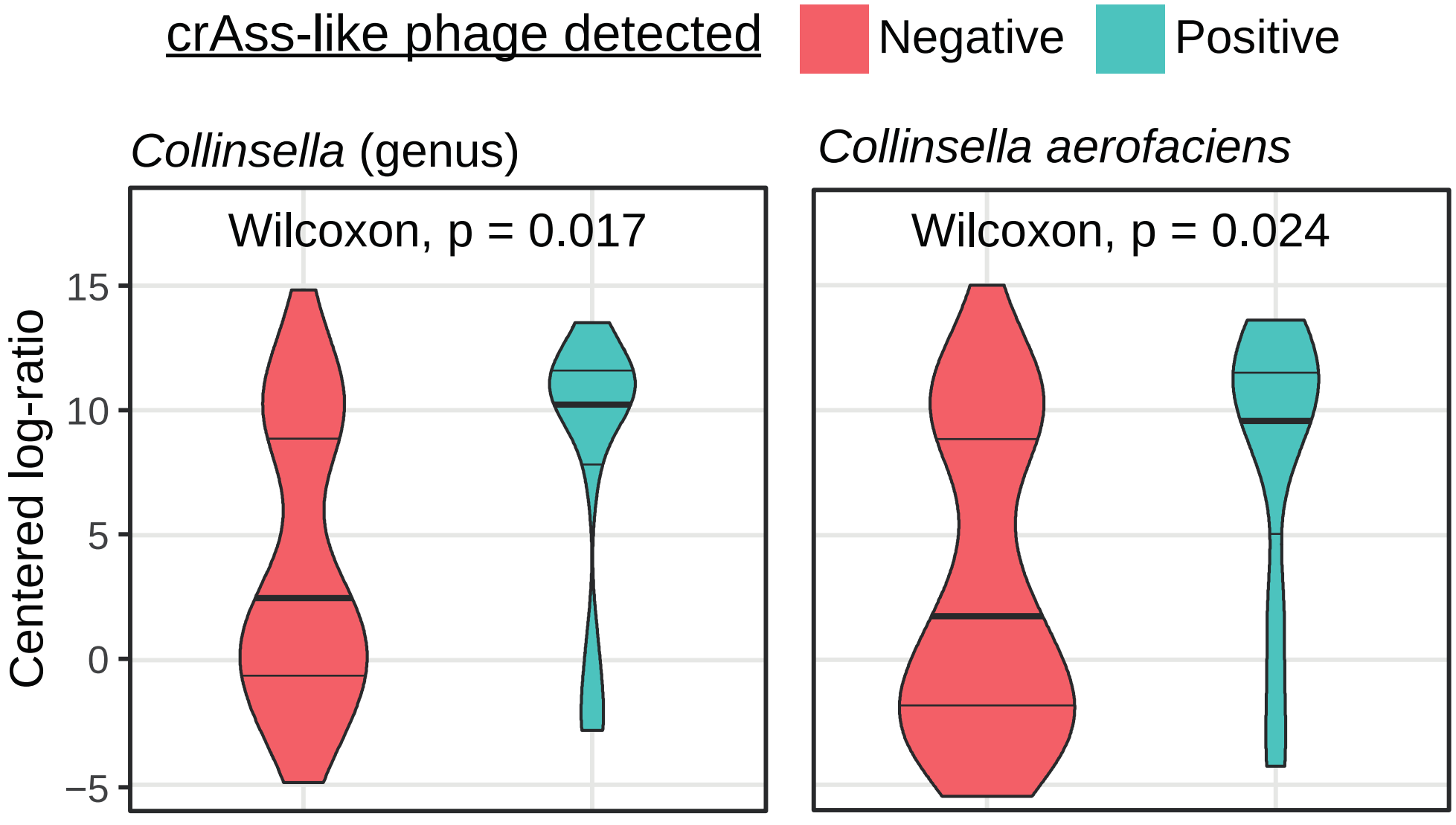

B
Yassour et al.
Bäckhed et al.
P-crAssphage positive
25%
20%
15%
10%
5%
0%
-3
0
3
0
Time since birth (months)

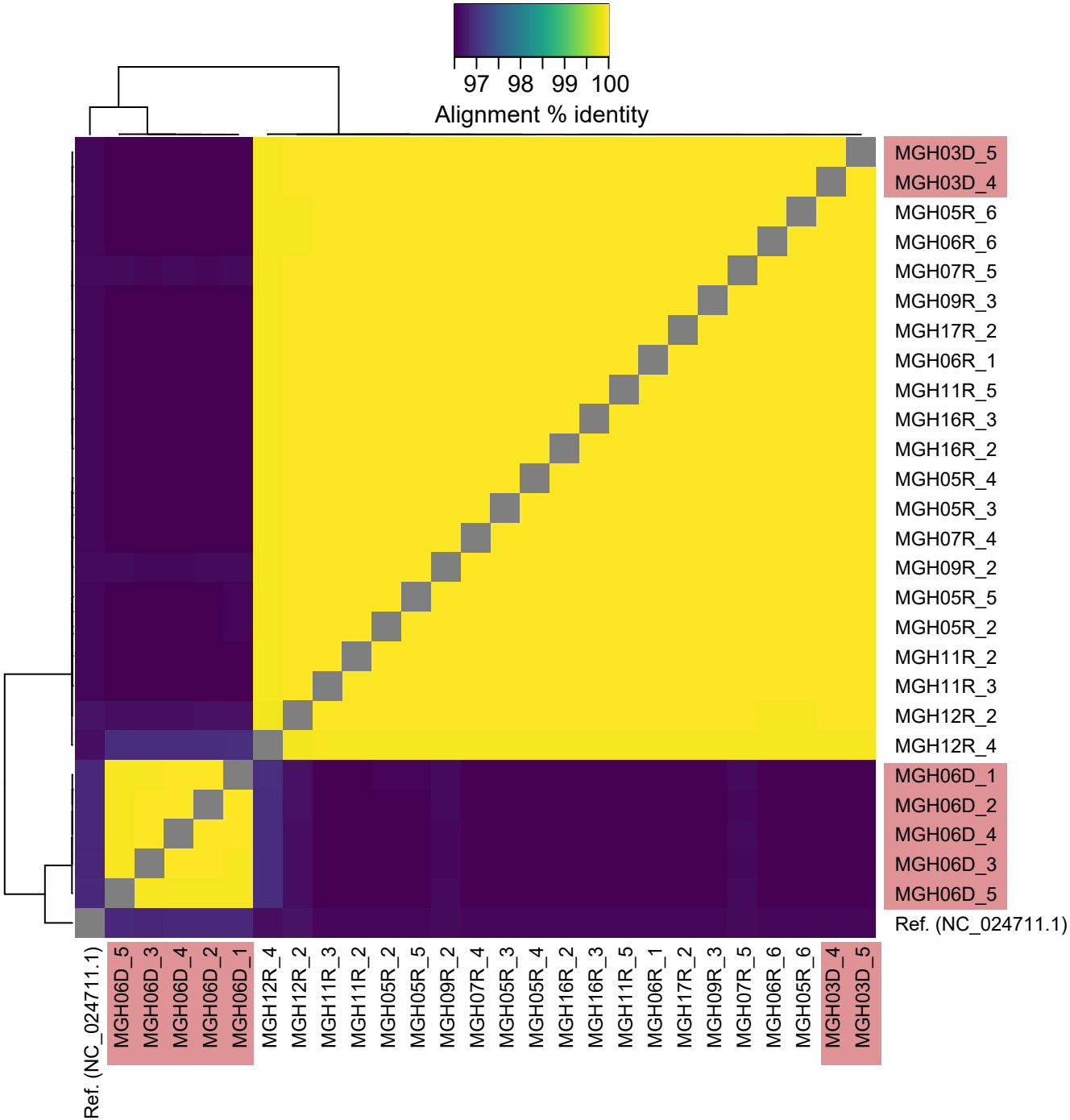

A
0.0016
p < 2.22e−16
1.9e−15
100
96
92
Alignment % identity
different family
same family
same individual
Comparison type
B
0.0022
p < 2.22e−16
1.9e−15
100
96
92
SNP % identity
different family
same family
same individual
Comparison type

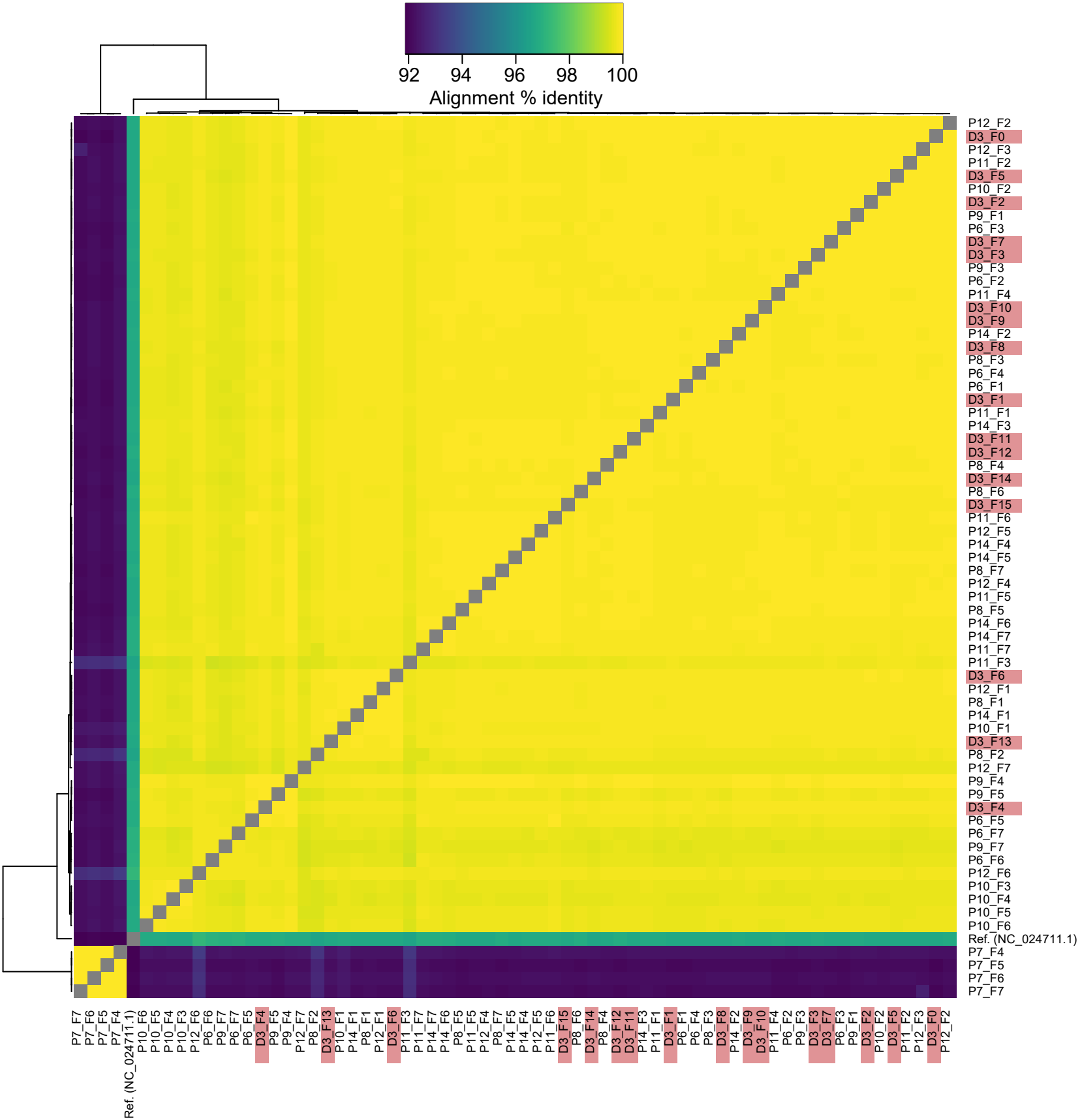

Alignment % identity
92
96
100
Family
M0226
345
268
272
335
M1098
M0808
263
343
184
272 Infant 12M
326 Mother birth
526 Mother birth
241 Infant 12M
511 Mother birth
Ref. (NC_024711.1)
274 Mother birth
78 Mother birth
256 Mother birth
M0388 Mother -3M
585 Mother birth
137 Mother birth
626 Mother birth
184 Infant 12M
203 Infant 12M
53 Infant 12M
53 Infant 4M
566 Mother birth
229 Infant 12M
624 Mother birth
635 Infant 12M
272 Mother birth
345 Mother birth
345 Infant 12M
70 Infant 12M
343 Infant 12M
343 Mother birth
332 Infant 12M
275 Mother birth
M0297 Mother birth
M0297 Mother -3M
M0297 Mother 3M
335 Infant 12M
335 Mother birth
615 Mother birth
549 Mother birth
387 Infant 4M
387 Infant 12M
263 Infant 12M
263 Infant 4M
263 Mother birth
184 Mother birth
383 Infant 12M
12 Mother birth
M0808 Mother birth
M0808 Mother -3M
M0808 Infant 3M
M0333 Mother birth
M0333 Mother 3M
M0333 Mother -3M
589 Mother birth
M0327 Mother 3M
M0327 Mother -3M
M0327 Mother birth
268 Mother birth
268 Infant 12M
M1098 Mother -3M
M1098 Mother 3M
M1098 Mother birth
M1098 Infant 2M
M1098 Infant 1M
M1098 Infant 3M
M0226 Mother -3M
M0226 Mother birth
M0226 Infant 2M
M0226 Infant 3M
M0226 Mother 3M
577 Mother birth
224 Mother birth
M0806 Mother -3M
M0806 Mother 3M
M0806 Mother 3M
M0806 Mother -3M
224 Mother birth
577 Mother birth
M0226 Mother 3M
M0226 Infant 3M
M0226 Infant 2M
M0226 Mother -3M
M1098 Infant 3M
M1098 Infant 1M
M1098 Infant 2M
M1098 Mother birth
M1098 Mother 3M
M1098 Mother -3M
268 Infant 12M
268 Mother birth
M0327 Mother birth
M0327 Mother -3M
M0327 Mother 3M
589 Mother birth
M0333 Mother -3M
M0333 Mother 3M
M0333 Mother birth
M0808 Infant 3M
M0808 Mother birth
M0808 Mother -3M
12 Mother birth
383 Infant 12M
184 Mother birth
263 Mother birth
263 Infant 4M
263 Infant 12M
387 Infant 12M
387 Infant 4M
549 Mother birth
615 Mother birth
335 Mother birth
335 Infant 12M
M0297 Mother 3M
M0297 Mother -3M
M0297 Mother birth
275 Mother birth
332 Infant 12M
343 Mother birth
343 Infant 12M
70 Infant 12M
345 Infant 12M
345 Mother birth
272 Mother birth
635 Infant 12M
624 Mother birth
229 Infant 12M
566 Mother birth
53 Infant 4M
53 Infant 12M
203 Infant 12M
184 Infant 12M
626 Mother birth
137 Mother birth
585 Mother birth
M0388 Mother -3M
256 Mother birth
78 Mother birth
274 Mother birth
Ref. (NC_024711.1)
511 Mother birth
241 Infant 12M
526 Mother birth
326 Mother birth
272 Infant 12M

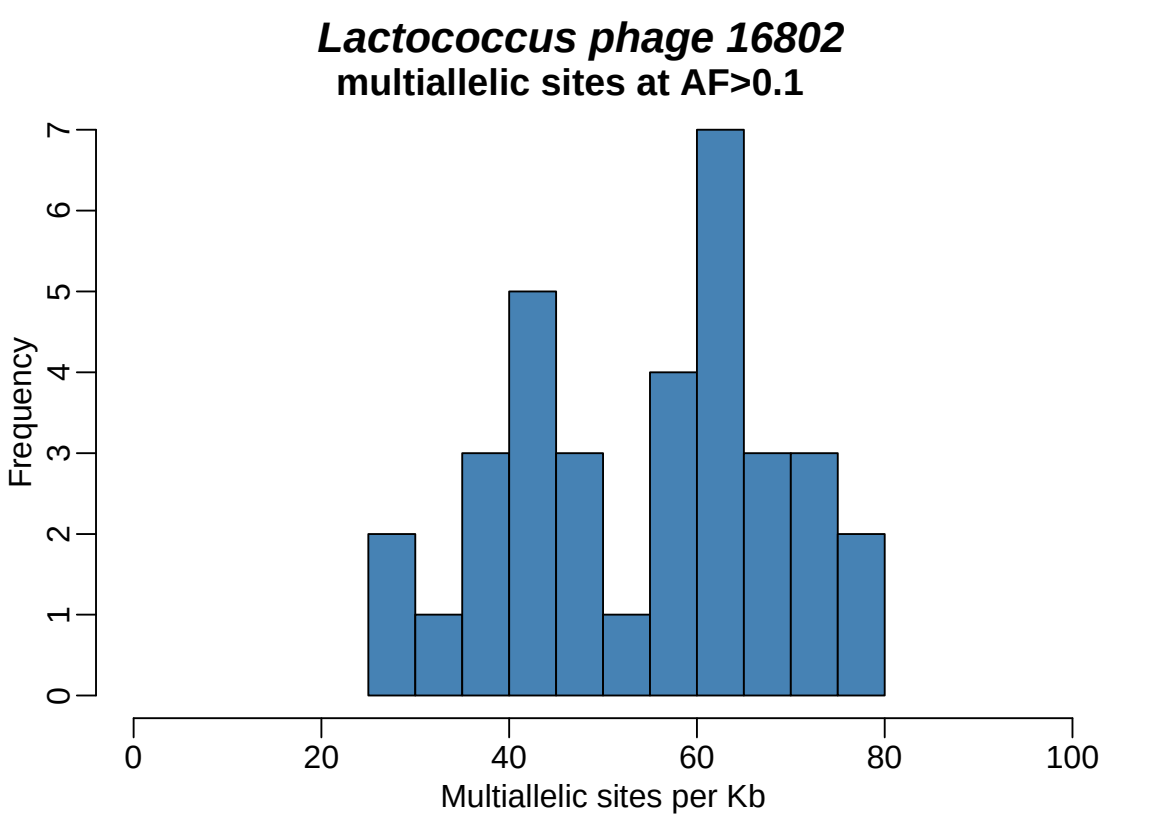

92
96
100
SNP % identity
Family
M0226
345
268
272
335
M1098
M0808
263
343
184
326 Mother birth
241 Infant 12M
526 Mother birth
585 Mother birth
275 Mother birth
184 Infant 12M
137 Mother birth
M0388 Mother -3M
256 Mother birth
229 Infant 12M
566 Mother birth
272 Infant 12M
511 Mother birth
Ref. (NC_024711.1)
274 Mother birth
53 Infant 12M
53 Infant 4M
549 Mother birth
272 Mother birth
635 Infant 12M
335 Mother birth
335 Infant 12M
343 Infant 12M
343 Mother birth
332 Infant 12M
387 Infant 12M
387 Infant 4M
263 Infant 12M
263 Infant 4M
263 Mother birth
184 Mother birth
383 Infant 12M
12 Mother birth
M0297 Mother birth
M0297 Mother -3M
M0297 Mother 3M
203 Infant 12M
70 Infant 12M
M0333 Mother -3M
M0333 Mother birth
M0333 Mother 3M
626 Mother birth
624 Mother birth
78 Mother birth
589 Mother birth
345 Mother birth
345 Infant 12M
M0808 Mother -3M
M0808 Mother birth
M0808 Infant 3M
M1098 Mother 3M
M1098 Mother birth
M1098 Mother -3M
M1098 Infant M
M1098 Infant 3M
M1098 Infant 2M
M0327 Mother -3M
M0327 Mother 3M
M0327 Mother birth
268 Mother birth
268 Infant 12M
M0226 Mother 3M
M0226 Mother birth
M0226 Infant 2M
M0226 Infant 3M
M0226 Mother -3M
M0806 Mother -3M
M0806 Mother 3M
577 Mother birth
224 Mother birth
224 Mother birth
577 Mother birth
M0806 Mother 3M
M0806 Mother -3M
M0226 Mother -3M
M0226 Infant 3M
M0226 Infant 2M
M0226 Mother birth
M0226 Mother 3M
268 Infant 12M
268 Mother birth
M0327 Mother birth
M0327 Mother 3M
M0327 Mother -3M
M1098 Infant 2M
M1098 Infant 3M
M1098 Infant M
M1098 Mother -3M
M1098 Mother birth
M1098 Mother 3M
M0808 Infant 3M
M0808 Mother birth
M0808 Mother -3M
345 Infant 12M
345 Mother birth
589 Mother birth
78 Mother birth
624 Mother birth
626 Mother birth
M0333 Mother 3M
M0333 Mother birth
M0333 Mother -3M
70 Infant 12M
203 Infant 12M
M0297 Mother 3M
M0297 Mother -3M
M0297 Mother birth
12 Mother birth
383 Infant 12M
184 Mother birth
263 Mother birth
263 Infant 4M
263 Infant 12M
387 Infant 4M
387 Infant 12M
332 Infant 12M
343 Mother birth
343 Infant 12M
335 Infant 12M
335 Mother birth
635 Infant 12M
272 Mother birth
549 Mother birth
53 Infant 4M
53 Infant 12M
274 Mother birth
Ref. (NC_024711.1)
511 Mother birth
272 Infant 12M
566 Mother birth
229 Infant 12M
256 Mother birth
M0388 Mother -3M
137 Mother birth
184 Infant 12M
275 Mother birth
585 Mother birth
526 Mother birth
241 Infant 12M
326 Mother birth


A
Months relative to birth
−3
0
3
Bäckhed et al. (2015) Mothers
Positive samples
75%
50%
25%
0%
Alpha 01
Alpha 03
Alpha 04
Alpha 09
Beta 06
Gamma 02
Gamma 05
Delta 07
Delta 08
Delta 10
any
B
Months relative to birth
<1
4
12
Bäckhed et al. (2015) Infants
Positive samples
40%
30%
20%
10%
0%
Alpha 01
Alpha 03
Alpha 04
Alpha 09
Beta 06
Gamma 02
Gamma 05
Delta 07
Delta 08
Delta 10
any
Yassour et al. (2018) Mothers
Positive samples
75%
50%
25%
0%
Alpha 01
Alpha 03
Alpha 04
Alpha 09
Beta 06
Gamma 02
Gamma 05
Delta 07
Delta 08
Delta 10
any
crAss−like phage cluster
Months relative to birth
0.5
1
2
3
Yassour et al. (2018) Infants
Positive samples
15%
10%
5%
0%
Alpha 01
Alpha 03
Alpha 04
Alpha 09
Beta 06
Gamma 02
Gamma 05
Delta 07
Delta 08
Delta 10
any
crAss−like phage cluster
C
Mothers
Fraction of samples
30%
20%
10%
0%
0
1
2
3
4
5
6
7
8
Number of clusters
Infants
80%
60%
40%
20%
0%
0
1
2
3
4
5
6
7
8
Number of clusters


crAss-like phage detected
Negative
Positive
Collinsella (genus)
Wilcoxon, p = 0.017
Collinsella aerofaciens
Wilcoxon, p = 0.024
Centered log-ratio
15
10
5
0
−5


97
98
99
100
Alignment % identity
MGH03D_5
MGH03D_4
MGH05R_6
MGH06R_6
MGH07R_5
MGH09R_3
MGH17R_2
MGH06R_1
MGH11R_5
MGH16R_3
MGH16R_2
MGH05R_4
MGH05R_3
MGH07R_4
MGH09R_2
MGH05R_5
MGH05R_2
MGH11R_2
MGH11R_3
MGH12R_2
MGH12R_4
MGH06D_1
MGH06D_2
MGH06D_4
MGH06D_3
MGH06D_5
Ref. (NC_024711.1)
Ref. (NC_024711.1)
MGH06D_5
MGH06D_3
MGH06D_4
MGH06D_2
MGH06D_1
MGH12R_4
MGH12R_2
MGH11R_3
MGH11R_2
MGH05R_2
MGH05R_5
MGH09R_2
MGH07R_4
MGH05R_3
MGH05R_4
MGH16R_2
MGH16R_3
MGH11R_5
MGH06R_1
MGH17R_2
MGH09R_3
MGH07R_5
MGH06R_6
MGH05R_6
MGH03D_4
MGH03D_5


92
94
96
98
100
Alignment % identity
P12_F2
D3_F0
P12_F3
P11_F2
D3_F5
P10_F2
D3_F2
P9_F1
P6_F3
D3_F7
D3_F3
P9_F3
P6_F2
P11_F4
D3_F10
D3_F9
P14_F2
D3_F8
P8_F3
P6_F4
P6_F1
D3_F1
P11_F1
P14_F3
D3_F11
D3_F12
P8_F4
D3_F14
P8_F6
D3_F15
P11_F6
P12_F5
P14_F4
P14_F5
P8_F7
P12_F4
P11_F5
P8_F5
P14_F6
P14_F7
P11_F7
P11_F3
D3_F6
P12_F1
P8_F1
P14_F1
P10_F1
D3_F13
P8_F2
P12_F7
P9_F4
P9_F5
D3_F4
P6_F5
P6_F7
P9_F7
P6_F6
P12_F6
P10_F3
P10_F4
P10_F5
P10_F6
Ref. (NC_024711.1)
P7_F4
P7_F5
P7_F6
P7_F7
P7_F7
P7_F6
P7_F5
P7_F4
Ref. (NC_024711.1)
P10_F6
P10_F5
P10_F4
P10_F3
P12_F6
P6_F6
P9_F7
P6_F5
D3_F4
P9_F5
P9_F4
P12_F7
P8_F2
D3_F13
P10_F1
P14_F1
P8_F1
P12_F1
D3_F6
P11_F3
P11_F7
P14_F7
P14_F6
P8_F5
P11_F5
P12_F4
P8_F7
P14_F5
P14_F4
P12_F5
P11_F6
D3_F15
P8_F6
D3_F14
P8_F4
D3_F12
D3_F11
P14_F3
P11_F1
D3_F1
P6_F1
P6_F4
P8_F3
D3_F8
P14_F2
D3_F9
D3_F10
P11_F4
P6_F2
P9_F3
D3_F3
D3_F7
P6_F3
P9_F1
D3_F2
P10_F2
D3_F5
P11_F2
P12_F3
D3_F0
P12_F2


Lactococcus phage 16802
multiallelic sites at AF>0.1
Frequency
0
1
2
3
4
5
6
7
0
20
40
60
80
100
Multiallelic sites per Kb