Genetic Evidence for Selective Transfer of Microbes Between the International Space Station and an Astronaut

David C. Danko[1,2], Nitin Singh[3], Daniel J. Butler[2], Christopher Mozsary[2], Peng Jiang[4], Ali Keshavarzian[5], Mark Maienschein-Cline[4], George Chlipala[4], Ebrahim Afshinnekoo[2], Daniela Bezdan[2], Fran Garrett-Bakelman[2, 6, 7], Stefan J. Green[4], Fred W. Turek[8], Martha Hotz Vitaterna[8], Kasthuri Venkateswaran[3], and Christopher E. Mason[2,10,*]

[1]Tri-Institutional Computational Biology & Medicine Program, Cornell University, NY, USA
[2]Institute for Computational Biomedicine, Department of Physiology and Biophysics, Weill Cornell Medicine of Cornell University, NY, USA
[3]Biotechnology and Planetary Protection Group, Jet Propulsion Laboratory, California Institute of Technology, Pasadena, Los Angeles, CA
[4]University of Illinois at Chicago, Chicago, IL, USA
[5]Rush University Medical Center, Chicago, IL USA
[6]Department of Medicine, University of Virginia, Charlottesville, VA
[7]Department of Biochemistry and Molecular Genetics, University of Virginia, Charlottesville, VA
[8]Northwestern University, Evanston, IL, USA
[9]The Feil Family Brain and Mind Research Institute, Weill Cornell Medicine, NY, USA
*Corresponding author

Abstract

Microbial transfer of both pathogenic and non-pathogenic strains from the environment can influence a person's health, but such studies are rare and the phenomenon is difficult to study. Here, we use the unique, isolated environment of the International Space Station (ISS) to track environmental movement of microbes in an astronaut's body. We identified several microbial taxa, including *Serratia proteamaculans* and *Rickettsia australis*, which appear to have been transferred from the environment of to the gut and oral microbiomes of the on-board astronaut, and also observed an exchange of genetic elements between the microbial species. Strains were matched at the SNP and haplotype-level, and notably some strains persisted even after the astronaut's return to Earth. Finally, some transferred taxa correspond to secondary strains in the ISS environment, suggesting that this process may be mediated by evolutionary selection, and thus, continual microbial monitoring can be important to future spaceflight mission planning and habitat design.

1 Introduction

Human commensal microbiomes have a known hereditary component (Goodrich et al., 2016), but the non-hereditary, acquired portion of the human microbiome is mediated by a large number of factors. An ideal study for microbial transfer would utilize a longitudinal sampling of subjects in a hermetically-sealed environment that was already profiled with strain-level resolution. The microbiome can change as a function of age, developmental stage, environmental exposures, antibiotic use, diet, and lifestyle, yet strain-level mapping and longitudinal tracking of such dynamics are limited. In particular, the movement of non-pathogenic microbes and how they can colonize an adult commensal microbiomes in a defined, quantified, and hermetically-sealed environment, is almost completely unknown (Schwendner et al., 2017).

Evidence for the transfer environmental microbes into adult commensal microbiomes could have important health implications, as it would provide a mechanism for how regional environmental microbiomes impact a person's microbiome. Cities in particular are known to host diverse environmental microbiomes (Danko et al., 2019) and transfer between commensal and environmental microbiomes may add to explanations for health differences between otherwise similar regions (Nicolaou et al., 2005). The selective transfer of certain microbial strains may also carry evolutionary implications for the microbes being transferred. If a microbial species can be shown to follow distinct selective patterns inside and outside of human commensal microbiomes it is possible that these patterns would eventually lead to strain or even species differentiation.

The ISS presents several advantages for the study of microbial transfer. As an environment, the ISS is well studied (as are its occupants), it is a uniquely sealed environment with essentially no chance of infiltration by exterior microbes between regular supply missions, and microgravity may lead to a general diffusion of microorganisms not present in more ordinary environments. Here, we present evidence for the transfer of environmental strains to an adult's gut and oral microbiome while on the International Space Station (ISS), during an almost year-long mission(Garrett-Bakelman et al., 2019). Of note, several of these strains were continuously observed after the mission, providing evidence of a persistent influence on the astronaut's microbiome, which may help to inform future studies on human microbial interaction.

2 Results

We collected 18 fecal and 23 oral microbiome samples from two identical twin human astronauts, one flight subject (TW, 9 stool, 6 saliva, 5 buccal) and one control who did not leave earth (HR, 9 stool, 7 saliva, 5 buccal), taken from 2014-2018. These were compared to 42 time-matched, environmental samples from the ISS that corresponded to the flight subject's mission duration. All samples were sequenced with 2x150bp read length to a mean depth of 12-15M reads (12.01, 14.96, and 14.97M mean reads for ISS, fecal, and saliva, respectively), then aligned to the catalog of NCBI RefSeq complete microbial genomes, examined for single nucleotide polymorphisms (SNPs), and then run with strain analysis with the MetaSUB CAP pipeline and Aldex2 (see methods).

2.1 Taxonomic profiles show evidence of continual microbial exchange

New taxa in flight subject (TW) match environmental and commensal microbiomes We first examined the proportion of taxa observed in a given sample that were not observed in a previous sample from the same donor. Any newly observed taxa in sample of a given type (e.g. stool) was annotated relative its presence in samples from other body or environmental sites (e.g saliva). For fecal samples, we segmented the previously unobserved taxa from each sample into four groups: taxa observed in any saliva sample taken before the given fecal samples, taxa observed in ISS samples but not observed in saliva, taxa observed in both ISS and saliva samples, and taxa that were not observed in either the ISS or the saliva. The same process was repeated for saliva samples but swapping fecal and saliva in the hierarchy. As expected,the time series of samples taken from the flight subject (TW) and ground control subject (HR) showed that earlier samples exhibited a greater proportion of novel organisms (Figure 1, S1).

Of note, each sample contains a number of unobserved taxa that matched taxa from saliva/feces or the ISS (even before flight), indicating these are common commensal species on Earth or possibly organisms absorbed in previous missions. Indeed both astronauts had previously been in the space station across multiple missions though with a 10-fold difference in duration (TW has logged 520 total days on the ISS vs. 54 days for HR). Interestingly, when we examined the fraction of taxa that match ISS taxa in pre-flight samples from TW compared to other samples from HR, a higher average rate (56%) of ISS-matching taxa was observed in pre-flight samples for TW relative to HR (51%), although not significant (p-value = 0.21). The fraction of taxa that matched different environments are listed in Table 1. For both saliva and fecal microbiomes the large majority of taxa at each time point had already been observed in a previous sample from that site.

A small number of taxa were never observed in any pre-flight sample from any body site but were observed in peri- and post-flight samples from TW. We filtered for taxa that had no reads observed in pre-flight samples and had at least ten reads in at least two peri- or post-flight samples. These taxa were further filtered for taxa that were observed in at least two ISS samples. The resulting list included five taxa: two viral genera, two viral species (both phage), and one bacterial species: *Rickettsia australis* (Figure 2). Given the generally low abundance of these taxa we cannot definitively rule out that they were present at an undetectable low threshold pre-flight. For comparison only 2 taxa (both viruses) met the above requirements in TW but were not identified in ISS samples.

Emergence of new taxa in gut microbiomes exceeds repeated sampling To place these taxonomic trends in context, we investigated whether the sampling time series from TW and HR would identify more new taxa than repeated assays on an unchanging fecal sample. We compared the fecal microbiome time series of TW and HR to 243 repeated samples taken from a single fecal sample (Sasada

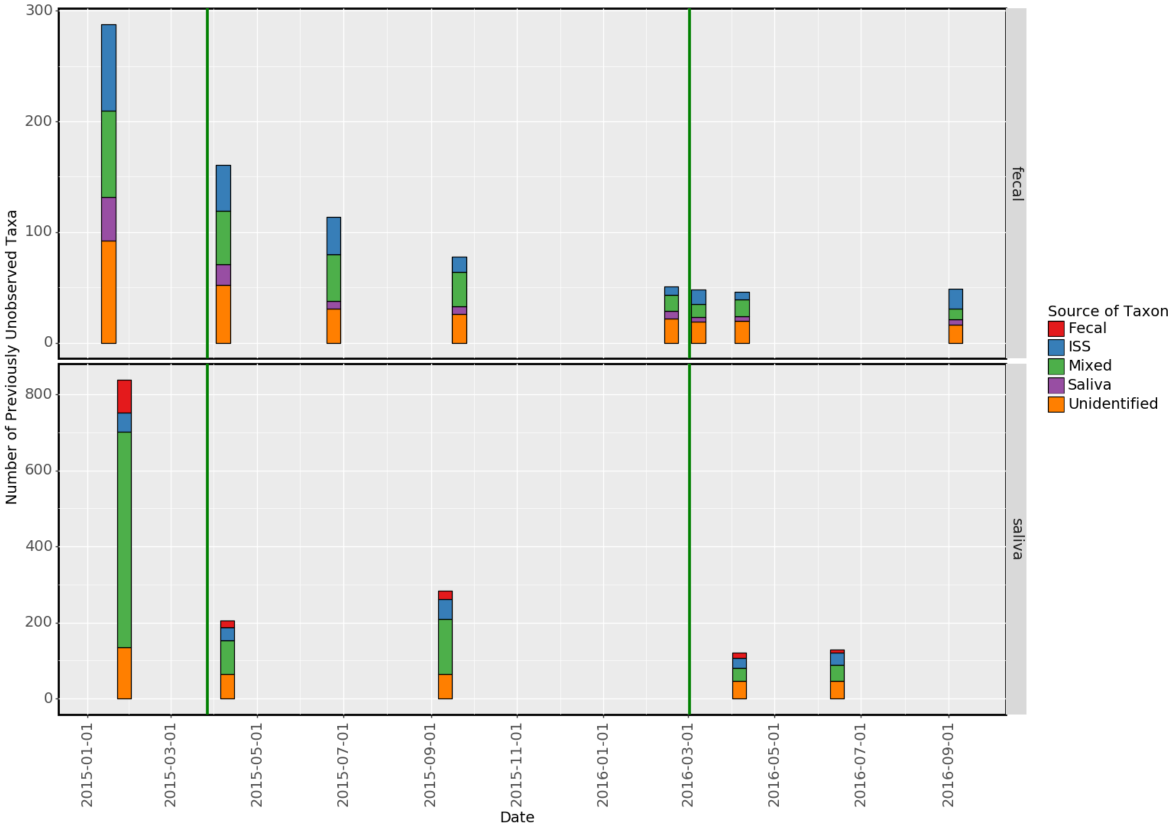

Figure 1: This plot shows the number of taxa at each time point that were not observed at any previous timepoint for fecal and saliva samples from TW. The colors indicate the likely source of the new taxon if it was found previously in the saliva (for fecal samples, vice versa for saliva samples), the ISS, both (Mixed), or neither.

Table 1: This table gives the average overlap between emergent taxa in fecal and saliva microbiomes and microbiomes in other sites.

Commensal Type	Fecal	Saliva
Sites Where Taxa Originated		
Fecal Only	n/a	8.7
Saliva Only	10.2	n/a
ISS Only	24.5	17.6
Both ISS & Saliva/Fecal	29.9	44.6
Taxa not identified in another site	35.5	29.1

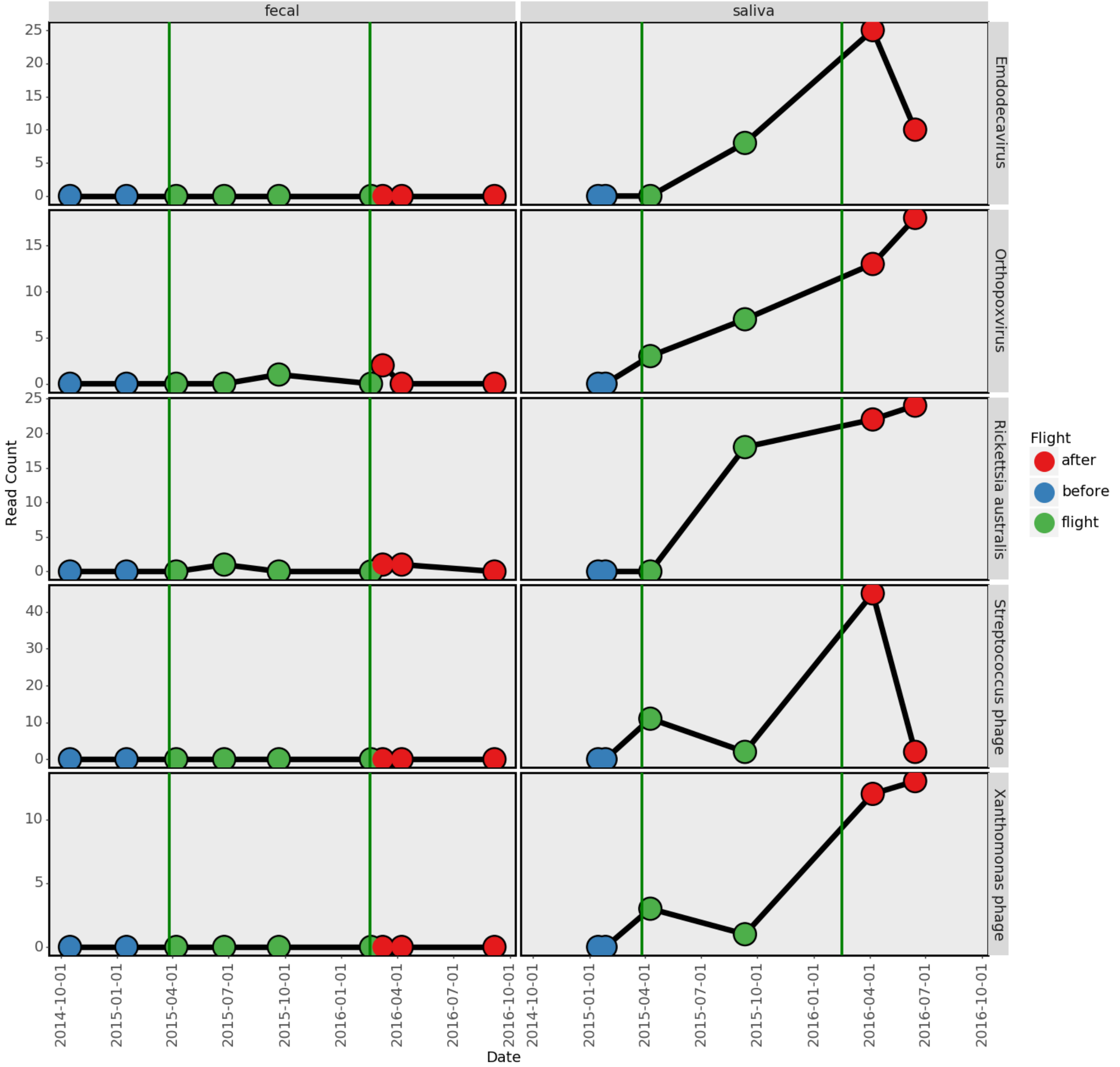

Figure 2: Total number of reads observed in TW for different taxa not observed before flight. Green vertical bars indicate the start and end of flight. The *Streptococcus phage* referenced is *phiARI0004*, *Xanthomonas phage* is *vB XveM DIBBI*,

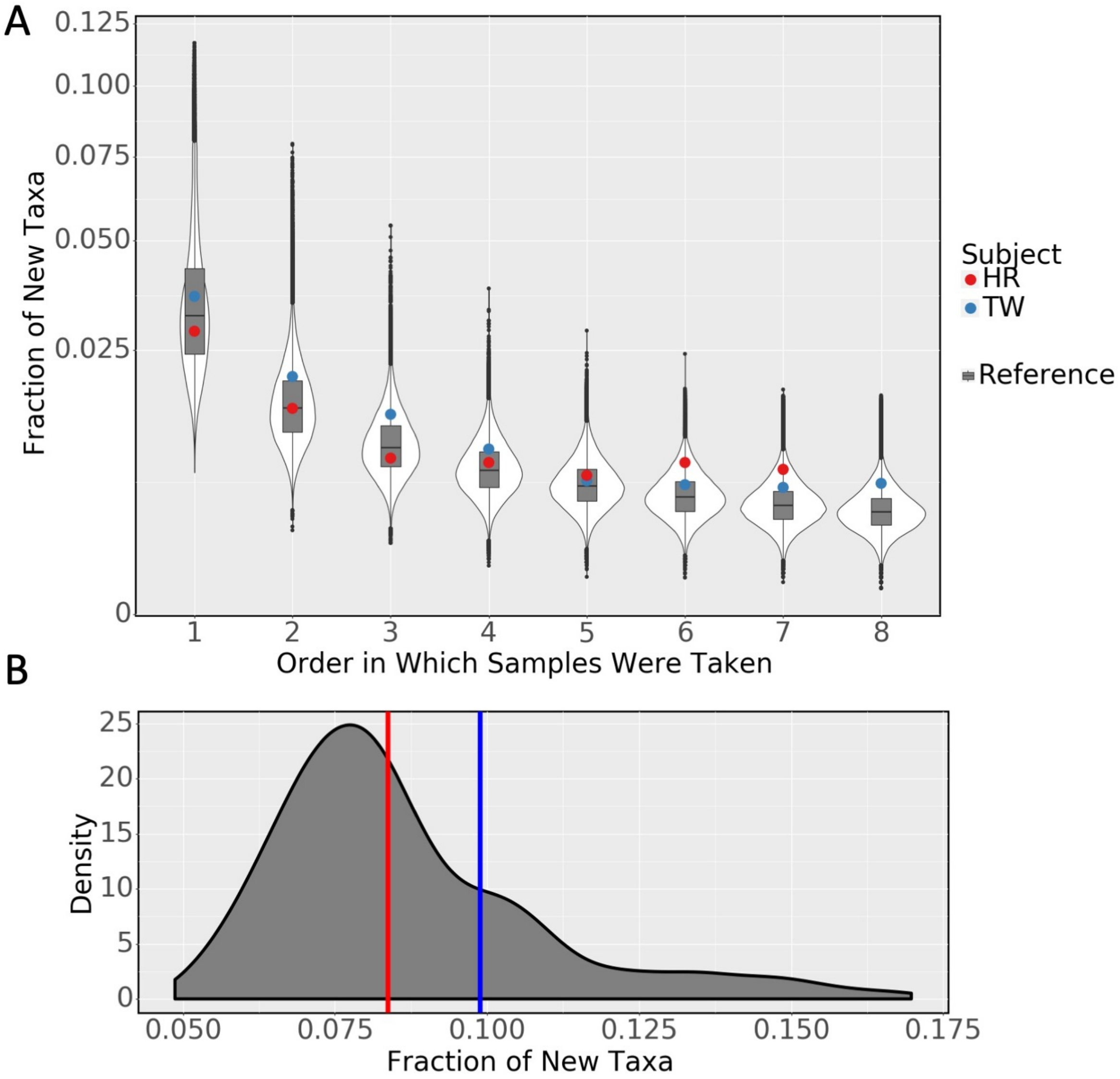

Figure 3: A) The number of new taxa observed in TW and HR are higher than repeated resampling of the same fecal sample. The y-axis gives the number of new taxa at each time point (not observed at any previous time point) divided by the number of taxa in the first sample. The first time point is omitted from the plot because it is always 1 by construction. The x-axis gives the order of each sample (arbitrary for random subsample). Boxplots show the distribution of random subsamples. Colored points are the actual time series. B) The number of unique taxa observed after the first time point divided by the number of taxa at the first time point. Same legend as (A)

et al., 2020), using 100,000 random sets of 9 samples (to match the twins' set size). The number of taxa in each sample that had not been observed in any previous sample were counted for each subset and normalized by the total number of taxa in the first sample. The time series for HR showed significantly more new taxa than 99,971 (p = 2.9e-4) random stool subsets, and TW more than 99,990 (p = 1e-4) random subsets (Figure 3). These results shows that the time series for TW and HR both consistently had more taxa than would be expected from resampling an unchanged fecal sample.

We note that differences between TW and HR may have a large number of causes including diet, environment, and exposure to other people. For this analysis it is relevant that both TW and HR show more new taxa than resampling a single sample since it implies transmission may also occur on earth.

Evidence of higher transfer rates on board the ISS We next calculated taxonomic diversity using Shannon's entropy for species profiles of each sample (Figure S2). For both fecal and saliva samples from TW, the highest diversity was observed during flight, and this trend was not observed for HR in the same time intervals. However, given the small sample size this trend was not significant (p=0.21).

We identified a significant increase in the number of previously unobserved taxa for samples taken from TW during flight (Figure S3) compared to random permutations. To further characterize the significance of such transfer relative to the sampling set and the source, we performed a series of permutation tests. We first established the number of previously unobserved species found at each time point in the actual data from TW. We then randomly shuffled and relabeled these samples and counted species again for a total of 10,000 random permutations. We then counted the number of permutations where the number of species observed 'during flight' in the shuffled data was higher than the real data . For the fecal microbiome the actual number of observed taxa was higher than the shuffled data in 96.7% of cases, for saliva 98.2% of cases and for buccal the observed was higher than all other permutations. Repeating the same procedure on data from HR ('flight' status was arbitrarily assigned to the second, third, and fourth samples) we observed 45.9% for feces 98.2% for saliva, and 80.1% for buccal (more buccal samples were available for HR).

Results were similar when the above procedure was repeated only with taxa found in ISS environmental samples (TW fecal 97.5%, TW saliva 97.7%, TW buccal all permutations, HR fecal 33.6%, HR saliva 98.3%, HR buccal 81.1%). An analogous analysis performed on ISS samples (Figure S4) showed that real data during TW's flight did not have significantly more new taxa than shuffled time periods (higher than 557 of 1,000 permutations). This is expected since the ISS is under continual habitation and merely is meant to show the converse of HR as a second control.

2.2 Strain level variation confirms microbial transfer

Novel genome regions in flight found in environmental and commensal microbiomes Given the higher overall transfer rate of species on the ISS, we next examined the strain emergence and persistence (post-flight) of such species. We selected a set of candidate taxa that showed significantly greater abundance during and after flight in TW than before flight. We mapped reads to known reference genomes from these taxa. We looked at the coverage of reference genomes at each stage of flight (concatenating samples from the same stage) and in the ISS and grouped regions into three categories: regions which were covered before flight, regions that were covered before flight in either gut or saliva samples but not observed in the other until flight, and regions that were not observed in either gut or saliva samples until flight but were found in the environment. Example coverage plots are shown for two taxa: *Fusobacterium necrophorum* and *Serratia proteamaculans* (Figure S5 and Figure 6 respectively). The total size of these genomic regions for all tested taxa are listed in Table 2.

For the selected taxa, the average environmental transfer of genomic regions were 32.2% of the size of pre-flight regions, whereas gut-saliva transfers were lower at 19.9%. The taxa with the (proportionally) largest transferred regions *Cronobacter condimenti*, had 55.9% gut-saliva transfer and 123.7% environmental transfer. The presence of (in some taxa) large genomic regions that were not covered until flight strongly suggests that individual species are undergoing flux with new strains and genes migrating into commensal microbiomes.

Microbial SNPs match environmental and commensal microbiomes Once the candidate genomic regions were identified, we next mapped co-occurring clusters of SNPs (haplotypes) in the selected taxa listed above in all samples from TW, HR, and the ISS (Figure 4). We matched microbial haplotypes from TW during flight to possible sources in pre-flight TW samples and ISS samples. Pre-flight fecal samples we considered four groups: haplotypes found in pre-flight fecal samples, haplotypes found

Table 2: Size of regions that may have been transferred in kilobases. Gut-Saliva transfer means that a region was found in either the gut or saliva microbiome pre-flight, then found in the other during-flight. Environment transfer means a region was not found in either fecal or saliva microbiomes from TW pre-flight but was found during flight and was also present in the ISS.

	Pre-flight	Gut-Saliva transfer	Environment transfer
Bifidobacterium pseudocatenulatum	243.9	92.4	85.2
Brevibacterium siliguriense	18.7	2.6	3.1
Gordonibacter urolithinfaciens	37.8	12.9	21.2
Bacillus albus	87.6	7.0	14.1
Gluconobacter albidus	10.2	2.5	1.3
Fusobacterium necrophorum	86.4	18.0	56.8
Geobacillus stearothermophilus	73.5	13.7	13.8
Bifidobacterium catenulatum	258.8	17.5	40.3
Streptococcus viridans	2319.6	92.9	221.8
Vibrio alginolyticus	211.0	10.6	89.7
Staphylococcus sciuri	179.0	19.4	37.4
Pectobacterium parmentieri	269.0	22.7	56.7
Campylobacter lari	42.0	8.7	18.1
Atlantibacter hermannii	66.4	15.7	30.6
Bacillus tequilensis	57.4	6.0	8.7
Achromobacter ruhlandii	49.8	13.6	11.7
Serratia proteamaculans	70.0	11.2	6.7
Leptotrichia hongkongensis	115.2	0.7	21.5
Exiguobacterium antarcticum	21.5	4.4	6.2
Anoxybacillus amylolyticus	11.5	2.2	2.3
Kosakonia sacchari	65.4	16.0	30.8
Yersinia canariae	18.2	8.2	7.7
Providencia heimbachae	76.0	12.1	6.5
Spirochaeta perfilievii	2.7	0.4	0.7
Cronobacter condimenti	15.2	8.5	18.8
Brenneria rubrifaciens	13.2	5.7	7.3
Staphylococcus simiae	20.8	1.5	6.2

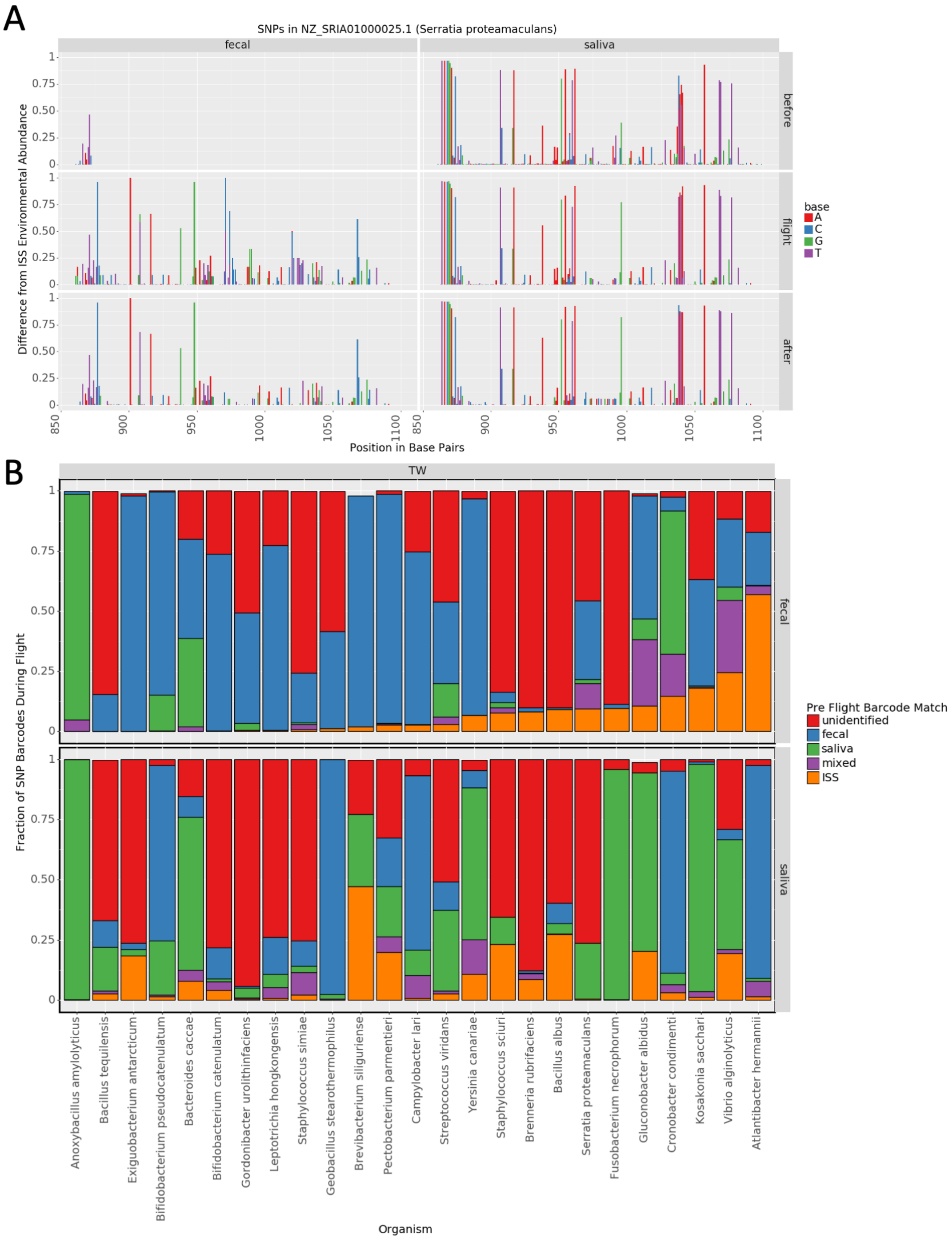

Figure 4: A) An example set of SNPs found in Serratia proteamaculans. The abundance of each SNP is shown relative to the frequency of the base found in the ISS at each position. A tall column indicates a base was low abundance in the ISS environment. In this case the SNPs shown for the fecal (left) strain match a secondary strain in the environment and constitute a candidate for transfer from the environment to the gut microbiome. B) Pre-flight sources of different SNP barcodes observed in TW during flight. Each SNP barcode in peri-flight samples from TW was matched to barcodes in pre-flight samples from TW and ISS samples. The fraction of barcodes matching each source is shown. For fecal samples barcodes labeled as saliva did not match fecal samples and vice versa. Barcodes labeled as matching ISS were not found in either fecal or saliva samples.

in pre-flight saliva but not fecal samples, haplotypes found in the ISS but neither saliva nor fecal, and haplotypes not observed in any other group.

The pre-flight sources of haplotypes varied by the species being investigated (Figure 4B). Some species, such as *Cronobacter condimenti* showed an apparent flip of strains from the gut microbiome to saliva and vice versa. Other taxa, like *Atlantibacter hermannii*, showed a large fraction of haplotypes that matched environmental haplotypes in the gut microbiome. Some taxa, like *Bifidobacterium catenulatum* showed little similarity to any potential external source.

2.3 Transfer case study: *Serratia proteamaculans*

***Serratia proteamaculans* (SP) is a candidate persistent transfer** We identified SP as a candidate persistent transfer, a species that was found in ISS environmental samples and was significantly more abundant in peri and post flight fecal samples from TW than in fecal samples from TW pre-flight and HR samples. As a whole SP was only found at low levels in fecal samples in TW pre-flight, was significantly more abundant during flight, and dropped to an intermediate level after flight (Figure 5). No major variation in abundance was observed for the control twin HR. SP was roughly uniformly abundant in the saliva before during and after flight.

Regions of the SP genome are found in TW fecal samples only after arrival at the ISS We identified regions of the SP genome which appeared in fecal samples after TW was on board the ISS. We found three such regions totaling about 1.5kbp. The abundance of these regions roughly matched the overall pattern seen for SP: very low or undetectable pre-flight, a high during flight, and an intermediate level post flight (Figure 6). These regions were all well covered from ISS environmental samples.

Total coverage of the SP genome in TW from all available fecal samples was 29.2kbp. Before flight 8.9kbp was covered, during 17.2kbp and after 19.0kbp. However some of these regions were either quite small or not covered in both peri and post flight. As such 1.5kbp represents a reasonable fraction of the amount of SP genome covered in TW but should only be interpreted as evidence for the transfer of particular genes.

2.4 SNPs in post-arrival regions match a secondary environmental strain

We analyzed one of the above regions (of about 250bp) for SNPs (Figure 4A) and identified SNPs in samples from TW which were either not found in the ISS environment or were found at different proportions. We identified 9 SNPs in this region during flight that were found in fewer than half of the ISS environmental samples. Of these 9 SNPs 6 were found after the conclusion of flight. We note that all of these 9 SNPs were found in ISS environmental samples at some proportion. We also note that this region did not match any other reference genome in RefSeq besides SP.

Next we used the SNP clustering technique described in the methods to determine if the 9 peri-flight SNPs we identified could come from the same strain. We identified corresponding groups of 8 SNPs in TW and 9 SNPs in the ISS environment. The 8 SNP group in TW included 8 out of the 9 peri-flight SNPs. The 9 SNP group from the ISS environment included these 8 SNPs as well as one SNP not identified in TW. This leads us to the conclusion that the strain found in TW likely represented a secondary strain in the ISS environment.

3 Methods

3.1 Experimental setup and samples

We analyzed 18 fecal samples from two human subjects (9 each) and 42 environmental samples from the ISS. All samples were assayed with 2x150bp DNA shotgun sequencing and analyzed as described below. Exact sample handling and processing is described in the supplementary methods.

Human fecal samples were taken from two identical twins TW and HR both astronauts who had previously been in space. During the study TW was sent on a roughly 1 year flight to the ISS while HR remained on earth and functioned as a control. For many parts of this study samples from TW are grouped into pre-flight, peri-flight, and post-flight groups. As much as practically possible samples from HR were handled in an identical manner to samples from TW.

We note that the sampling of the ISS was initially planned and designed separately from the sampling of the human subjects.

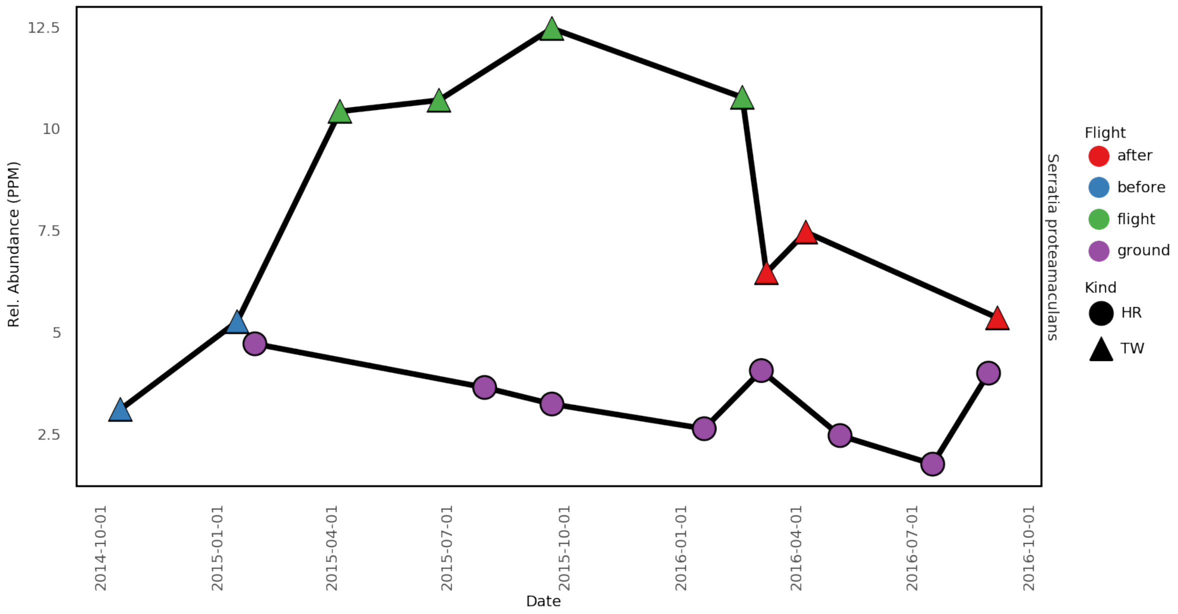

Figure 5: Relative abundance of *Serratia proteamaculans* in fecal samples from TW and HR. Relative abundance is given in units of parts per million.

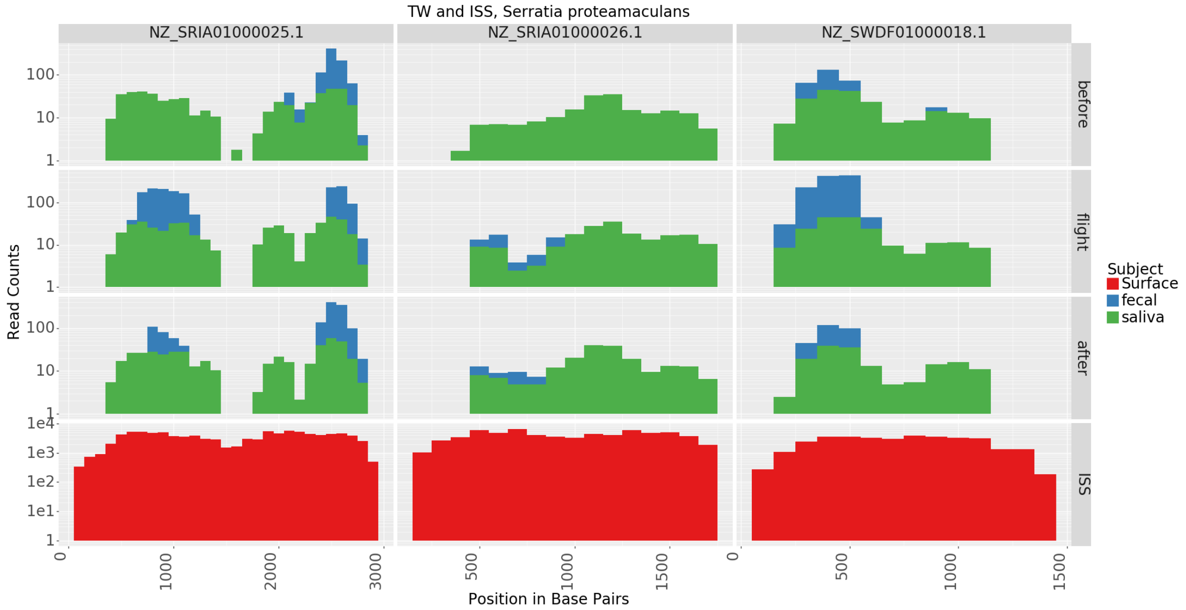

Figure 6: Coverage of candidate persistent transfer regions of the *Serratia proteamaculans* genome.

3.2 Sequencing

Samples from the human subject were extracted with a DNA extraction protocol adapted from the Maxwell RSC Buccal Swab DNA kit (Catalogue number AS1640: Promega Corporation, Madison WI). Briefly, 300 µl of lysis buffer and 30 µl of Proteinase K was mixed and added to each swab tube. Swab tubes were then incubated for 20 min at 56 C using a Thermo Fisher water bath, removed from the tubes, and fluid was transferred to well 1 of the Maxwell RSC Cartridge. The swab head was centrifuged using a ClickFit Microtube (Cat. # V4741), and extracted fluid was added to the corresponding well of Maxwell Cartridge, and eluted in 50 µl of provided elution buffer.

Extracted DNA was taken forward to the Nextera Flex protocol by Illumina. Briefly, 30 µl of extracted DNA was taken into library prep protocol and run with 12 cycles of PCR. Libraries were cleaned up with a left sided size selection, using a bead ratio of 0.8x. The right sided size selection was omitted. Libraries were then quantified using a Thermo Fisher Qubit Fluorometer and an Advanced Analytical Fragment Analyzer. Libraries were sequenced on an Illumina HiSeqPE 50×2 at the Weill Cornell Epigenomics Core.

Samples from the ISS were sequenced according to the protocol described in Singh et al. (2018).

3.3 Processing Short Read Sequencing Data

Preprocessing and Taxonomic Profiling We processed raw reads from all samples into taxonomic profiles for each sample using the MetaSUB Core Analysis Pipeline (Danko and Mason, 2020). This includes a preprocessing stage that consists of AdapterRemoval (Schubert et al., 2016), Human sequence removal with Bowtie2 (Langmead and Steven L Salzberg, 2013), and read error correction using BayesHammer (Nikolenko et al., 2013). Subsequently reads were assigned to taxonomic groups using Kraken2 (Wood et al., 2019). We generated a table of read counts giving the number of reads assigned to each species for each sample.

Identification of candidate species for strain level analysis We analyzed our table of species level read counts to identify candidate lists of *transient* and *persistent* transfer species. We held a transient species to be one that was transferred from the ISS into the astronaut only while the astronaut remained in the ISS and which was be cleared after return to earth. We held persistent species to be those that were transferred from the ISS to the astronaut which remained after return to earth.

We statistically analyzed our table of read counts using Aldex2 (Fernandes et al., 2013). Remaining samples (from astronauts) were split into two groups. The first group was the control group and consisted of all samples from TW before flight and all samples from HR at any point. The second group was the case group and consisted of all samples from TW during flight. Samples from TW after flight were assigned to the control group for analysis of transients and to the case group for analysis of persistents. Aldex2 was used to identify differentially abundant taxa between the two groups. We selected all taxa that were significantly ($q < 0.05$ by Welch's t-test with Benjamini Hochberg correction) more abundant in the case group than in the control group. We then filtered these two list (persistent and transient) to include only species found in the ISS samples (minimum 10 reads in 25% of samples).

Strain Analysis Reads were further processed for strain level analysis using the MetaSUB Core Analysis Pipeline. Given a specified organism to examine we downloaded all available reference genomes from RefSeq. If more than 100 reference genomes were available we selected 100 at random. Human-depleted reads were mapped to each genome using Bowtie2 (sensitive presets) and pileup files were generated using from alignments using samtools (Li et al., 2009). Pileups were analyzed for coverage patterns using purpose built code (see availability for access). SNPs were identified by comparing aligned bases from short reads to reference sequences, SNP filtering was performed as part of identifying co-stranded SNPs.

Identifying co-stranded SNPs We developed a technique to identify SNPs that occurred on the same genetic strand. The technique is, in practice, limited to identifying co-stranded SNPs within 1kbp of on another. The technique works by formulating SNP recovery as an instance of the multi-community recovery problem. We start by building a graph of SNPs. Each SNP forms a node in the graph and is identified by its genomic position and base. Edges are added between SNPs that are found on the same read. Edges are undirected but weighted by the number of times a pair of SNPs is found on the same read. The SNP graph is then filtered to remove SNPs that occur only once as these are likely to be errors and are uninformative in any case. The remaining graph is clustered into groups of SNPs using the approach to the multi-community recovery problem by Blondel et al. (2008). The final result of this are sets of SNPs that are often found on the same read.

This technique is similar to techniques used for phasing SNPs to one strand of a diploid genome such as Zheng et al. (2016). The key difference between this technique and ours is that there may be more than two communities in our case and that we make only attempt to cluster proximal SNPs.

4 Conclusion

We have identified genetic evidence of microbial transfer between the fecal and saliva microbiomes of an adult and between these microbiomes and their environment. These results demonstrate that non-pathogenic microbes from the environment can establish themselves in adults and suggests the possibility of ongoing microbial flux between humans and the unique ISS environment. Moreover, these provide candidate "ISS mobile" species and also enable a key estimate of the fraction of taxa that could be transferred from different sources of the body while in the spaceflight environment.

A number of open questions remain. We have made a first attempt to quantify the rate of transfer between different microbiomes and given an estimate for the total number of emergent species in a gut microbiome which cannot be explained as the result of repeated sampling alone. However, these estimates necessarily suffer from the small sample sizes available in this study and the unusual situation under which the samples were taken. To conclusively establish the scope of microbial transfer will require broader studies targeting earth based environments, food, and communities as well as confirmation using culture-based techniques. Nonetheless, the unusual nature of spaceflight provides as strongly controlled an environment as is likely to be possible making this a near-optimal model set up to study microbial transfer.

The emergence of new taxa, while intriguing, must be placed into the context of expected stool sampling variation. To account for such sampling dynamics, we also conducted a rigorous re-sampling study. Our data showed that TW and HR had more newly observed taxa at some (but not all) of the time points relative to the 100,000 subset. Importantly, the number of new taxa that were observed in subsets dropped off quickly for later time points as the subsets reached saturation. Subsets generally showed an adversarial selection, wherein many new taxa at one time point would lead to fewer new taxa at later time points. The 243 fecal replicates had similar read counts to the time series from HR and TW, reducing a source of potential bias, but could also be examined in greater detail in future studies.

Of note, repeated sampling can identify low abundance species which were dropped out of previous samples and because different sample preparation techniques can yield different sets of taxa. A series of samples taken from a microbiome that is exchanging taxa with an external environment will have an additional source of new taxa. These taxa would not be identified in earlier samples because they were not present, and this is another source of variation that could be mapped and quantified for future missions (more sampling of more areas of the body and the ISS, and at greater depth).

Taken together, the matching genomic regions across 16 taxa and matching SNPs haplotypes within the regions strongly supports the conclusion that novel taxa in pre-flight commensal microbiomes from TW could come from the environment or from other commensal microbiomes. The size of transferred regions and number of SNPs suggests that "taxa transfer" between commensal microbiomes occurs more frequently than they transfer from the environment to commensal microbiomes. However, these rates may prove to be anomalous for either TW, habitation in the ISS, or both, since non-pathogenic microbial exchange with the environment represents a significant unknown for its impact on human and astronaut health. Nevertheless, accurate quantification of microbial strains and their movements can lead to targeted interventions, shed light on the hygiene hypothesis (broadly and on the ISS), and help in planning for future missions and astronaut monitoring.

5 Availability and Access

All analysis and figure generating code may be found on GitHub at `https://github.com/dcdanko/twins_iss_transfer`. All results and raw data may be found on Pangea at `https://pangea.gimmebio.com/sample-groups/62661efb-a433-4ae5-bcec-de704a80e217`.

6 Author Contribution

DCD performed all bioinformatics analyses and defined the structure of the study. NS led the collection of samples from the ISS. DJB and CM prepared samples for sequencing. PJ, AK, MMC, GC, EA, coordinated sampling. FGB prepared samples for sequencing. SJG and MHV handled sample coordination, sequencing, collection, analysis. KV led coordination with NASA and led collection of samples on board the ISS. CEM led and conceived this study.

7 Acknowledgment

We would like to thank the Epigenomics Core Facility at Weill Cornell Medicine, the Scientific Computing Unit (SCU), XSEDE Supercomputing Resources, the Starr Cancer Consortium (I13-0052), the Vallee Foundation, the WorldQuant Foundation, The Pershing Square Sohn Cancer Research Alliance, NASA (NNX14AH50G, NNX17AB26G), the National Institutes of Health (R01AI151059), TRISH (NNX16AO69A:0107, NNX16AO69A:0061), the Bill and Melinda Gates Foundation (OPP1151054), and the Alfred P. Sloan Foundation (G-2015-13964).

Part of the research described in this article was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with NASA. This research was funded by 2012 Space Biology project NNH12ZTT001N grant 19-12829-26 under task order NNN13D111T awarded to K.V., which also funded a postdoctoral fellowship for NKS.

References

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008.

Danko, D., Bezdan, D., Afshinnekoo, E., Ahsanuddin, S., Bhattacharya, C., Butler, D. J., Chng, K. R., De Filippis, F., Hecht, J., Kahles, A., et al. (2019). Global genetic cartography of urban metagenomes and anti-microbial resistance. *BioRxiv*, page 724526.

Danko, D. C. and Mason, C. (2020). The metasub microbiome core analysis pipeline enables large scale metagenomic analysis.

Fernandes, A. D., Macklaim, J. M., Linn, T. G., Reid, G., and Gloor, G. B. (2013). ANOVA-Like Differential Expression (ALDEx) Analysis for Mixed Population RNA-Seq. *PLoS ONE*, 8(7).

Garrett-Bakelman, F. E., Darshi, M., Green, S. J., Gur, R. C., Lin, L., Macias, B. R., McKenna, M. J., Meydan, C., Mishra, T., Nasrini, J., et al. (2019). The nasa twins study: A multidimensional analysis of a year-long human spaceflight. *Science*, 364(6436).

Goodrich, J. K., Davenport, E. R., Beaumont, M., Jackson, M. A., Knight, R., Ober, C., Spector, T. D., Bell, J. T., Clark, A. G., and Ley, R. E. (2016). Genetic Determinants of the Gut Microbiome in UK Twins. *Cell Host and Microbe*, 19(5):731–743.

Langmead and Steven L Salzberg (2013). Bowtie2. *Nature methods*, 9(4):357–359.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079.

Nicolaou, N., Siddique, N., and Custovic, A. (2005). Allergic disease in urban and rural populations: Increasing prevalence with increasing urbanization. *Allergy: European Journal of Allergy and Clinical Immunology*, 60(11):1357–1360.

Nikolenko, S. I., Korobeynikov, A. I., and Alekseyev, M. A. (2013). BayesHammer: Bayesian clustering for error correction in single-cell sequencing. *BMC Genomics*, 14.

Sasada, R., Weinstein, M., Danko, D., Wolfe, E., Tang, S., Jarvis, K., Grim, C., Lagishetty, V., Jacobs, J., Arnold, J., Kemp, R., and Mason, C. (2020). Progress Towards Standardizing Metagenomics: Applying Metagenomic Reference Material to Develop Reproducible Microbial Lysis Methods with Minimum Bias. *Journal of biomolecular techniques : JBT*, 31:S30–S31.

Schubert, M., Lindgreen, S., and Orlando, L. (2016). AdapterRemoval v2: rapid adapter trimming, identification, and read merging. *BMC Research Notes*, 9(1):88.

Schwendner, P., Mahnert, A., Koskinen, K., Moissl-Eichinger, C., Barczyk, S., Wirth, R., Berg, G., and Rettberg, P. (2017). Preparing for the crewed Mars journey: microbiota dynamics in the confined Mars500 habitat during simulated Mars flight and landing. *Microbiome*, 5(1):129.

Singh, N. K., Wood, J. M., Karouia, F., and Venkateswaran, K. (2018). Succession and persistence of microbial communities and antimicrobial resistance genes associated with International Space Station environmental surfaces. *Microbiome*, 6(1).

Wood, D. E., Lu, J., and Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. *Genome Biology*, 20(1).

Zheng, G. X., Lau, B. T., Schnall-Levin, M., Jarosz, M., Bell, J. M., Hindson, C. M., Kyriazopoulou-Panagiotopoulou, S., Masquelier, D. A., Merrill, L., Terry, J. M., Mudivarti, P. A., Wyatt, P. W., Bharadwaj, R., Makarewicz, A. J., Li, Y., Belgrader, P., Price, A. D., Lowe, A. J., Marks, P., Vurens, G. M., Hardenbol, P., Montesclaros, L., Luo, M., Greenfield, L., Wong, A., Birch, D. E., Short, S. W., Bjornson, K. P., Patel, P., Hopmans, E. S., Wood, C., Kaur, S., Lockwood, G. K., Stafford, D., Delaney, J. P., Wu, I., Ordonez, H. S., Grimes, S. M., Greer, S., Lee, J. Y., Belhocine, K., Giorda, K. M., Heaton, W. H., McDermott, G. P., Bent, Z. W., Meschi, F., Kondov, N. O., Wilson, R., Bernate, J. A., Gauby, S., Kindwall, A., Bermejo, C., Fehr, A. N., Chan, A., Saxonov, S., Ness, K. D., Hindson, B. J., and Ji, H. P. (2016). Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nature Biotechnology*, 34(3):303–311.

Supplement

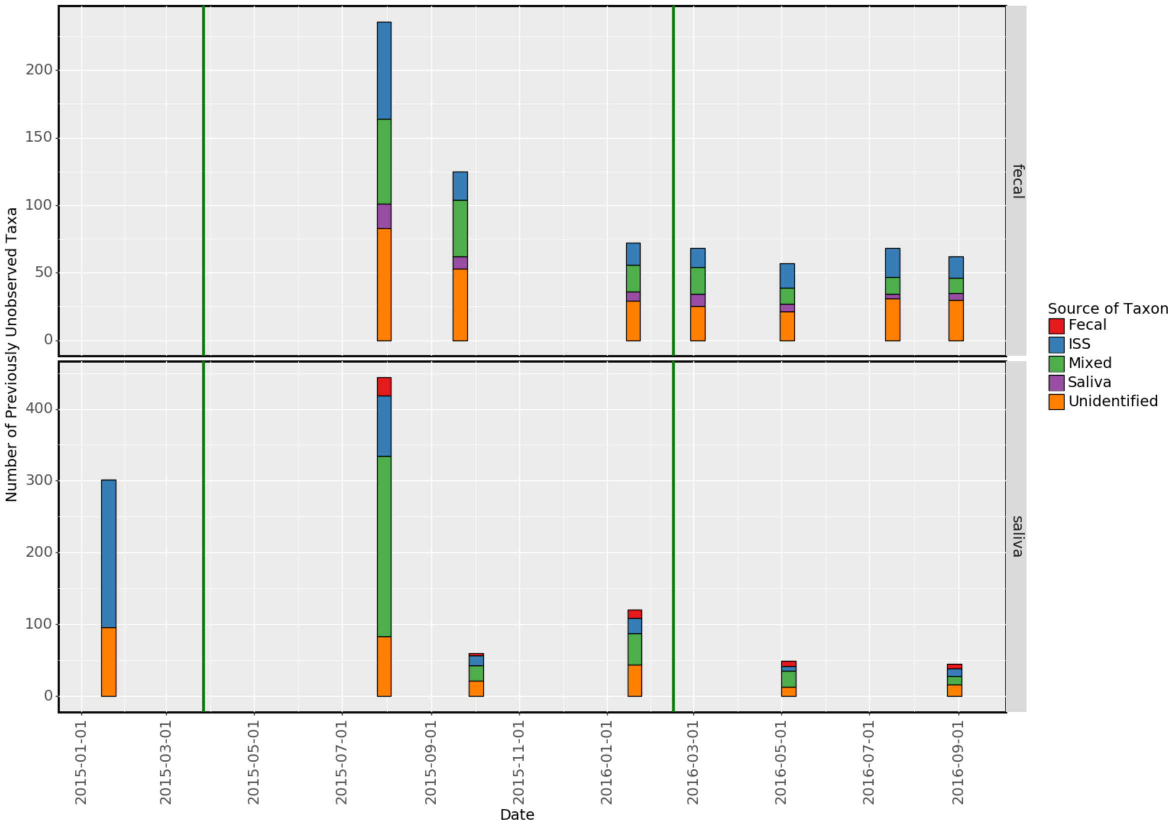

Figure S1: This plot shows the number of taxa at each time point that were not observed at any previous timepoint for fecal and saliva samples from HR. The colors indicate the likely source of the new taxon if it was found previously in the saliva (for fecal samples, vice versa for saliva samples), the ISS, both (Mixed), or neither.

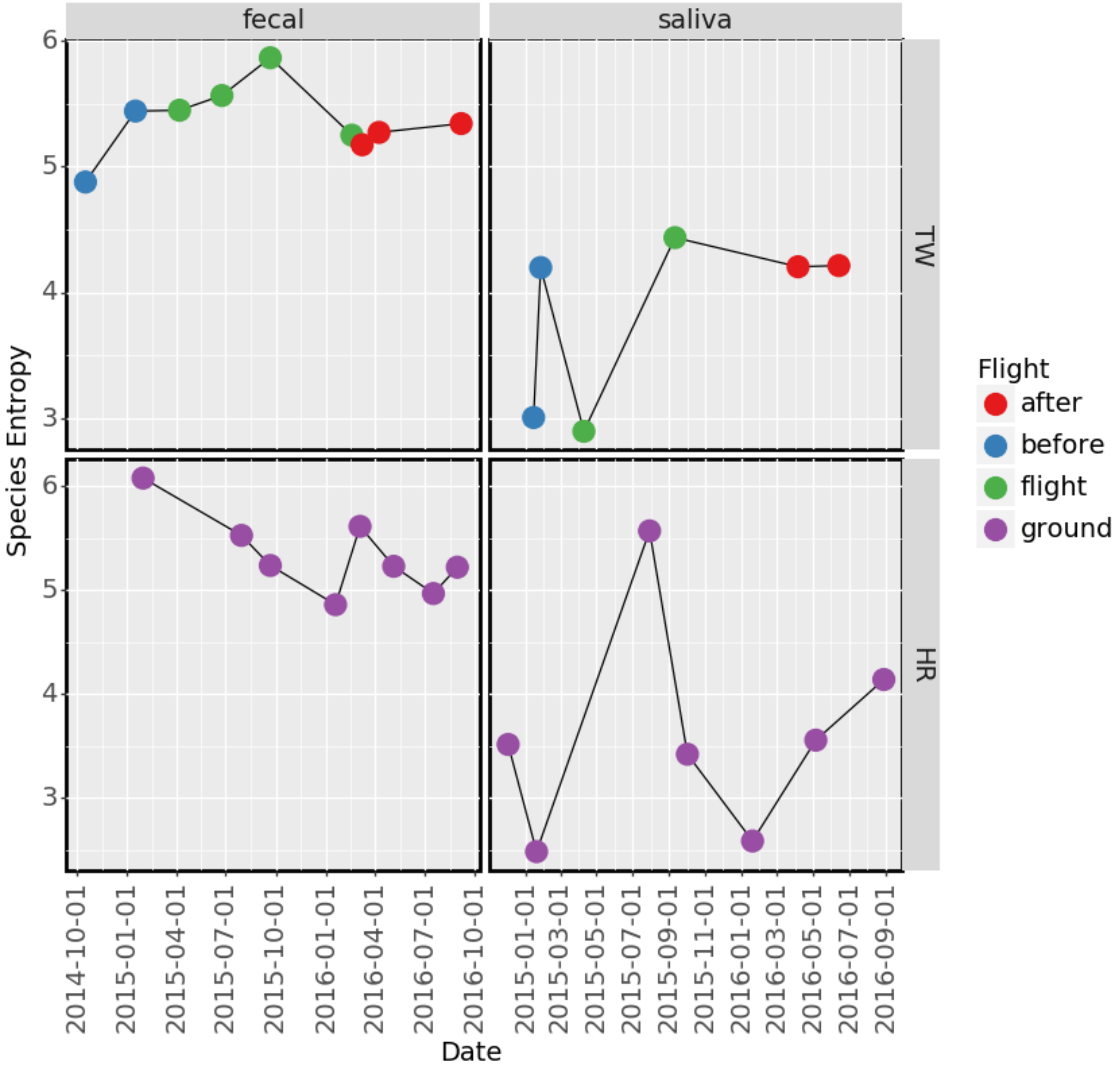

Figure S2: Vertical shows species entropy (Shannon entropy of species relative abundances) for sample types in both twins.

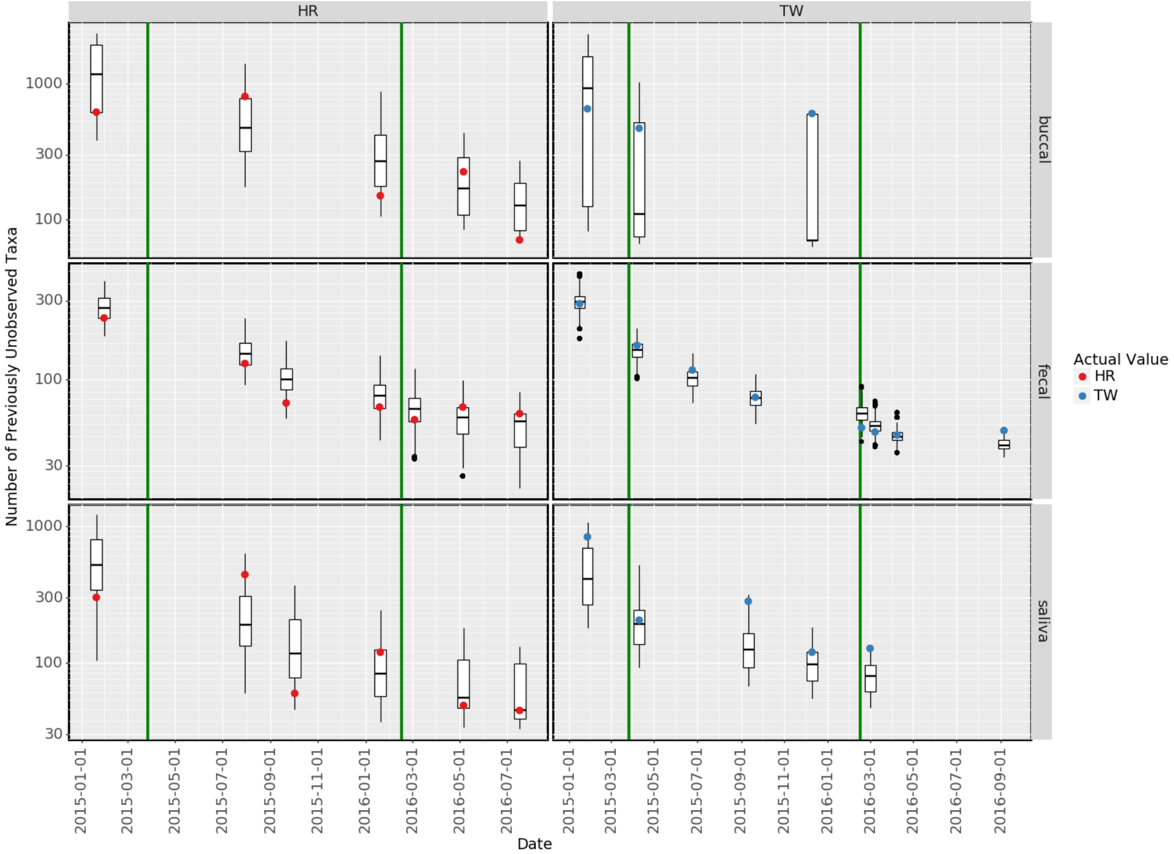

Figure S3: This plot shows the number of taxa at each time point that were not observed at any previous timepoint. The first timepoint is omitted from the plot since no taxa had been previously observed. Boxplots indicate an artificial reference distribution generated by randomly permuting timestamps. Red and blue dots indicate actual values.

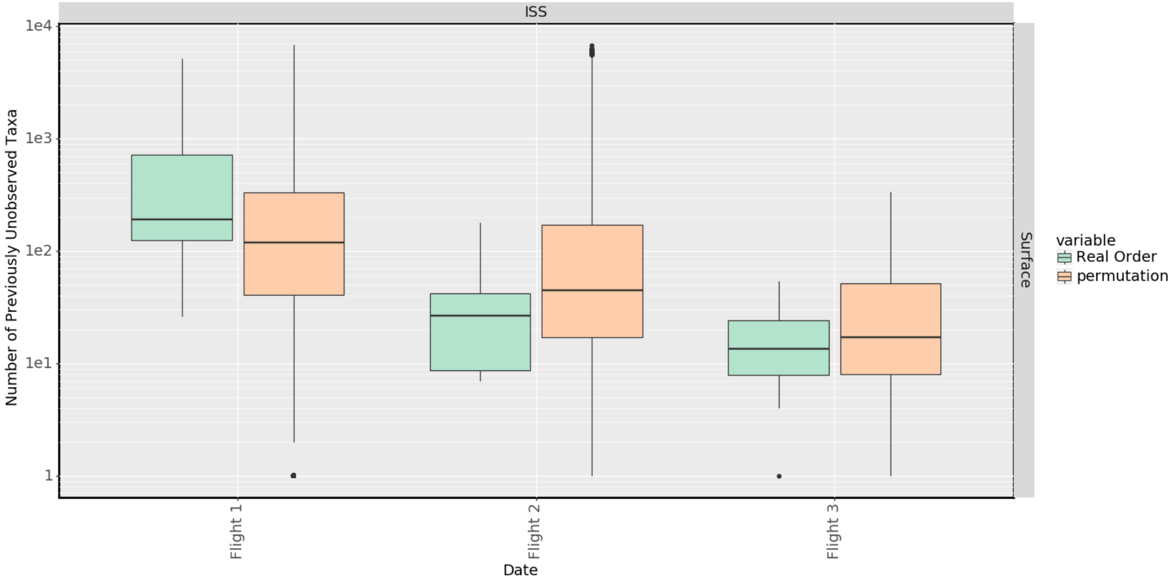

Figure S4: This plot shows the number of taxa at each time point that were not observed at any previous time point for the ISS. ISS samples are grouped into 'flights' where each sample in the same flight was taken on the same day. One sample from flight 1 is arbitrarily chose as the 'first' sample and used as the comparison. Boxplots indicate the real distribution of new taxa as well as an artificial reference distribution generated by randomly permuting timestamps.

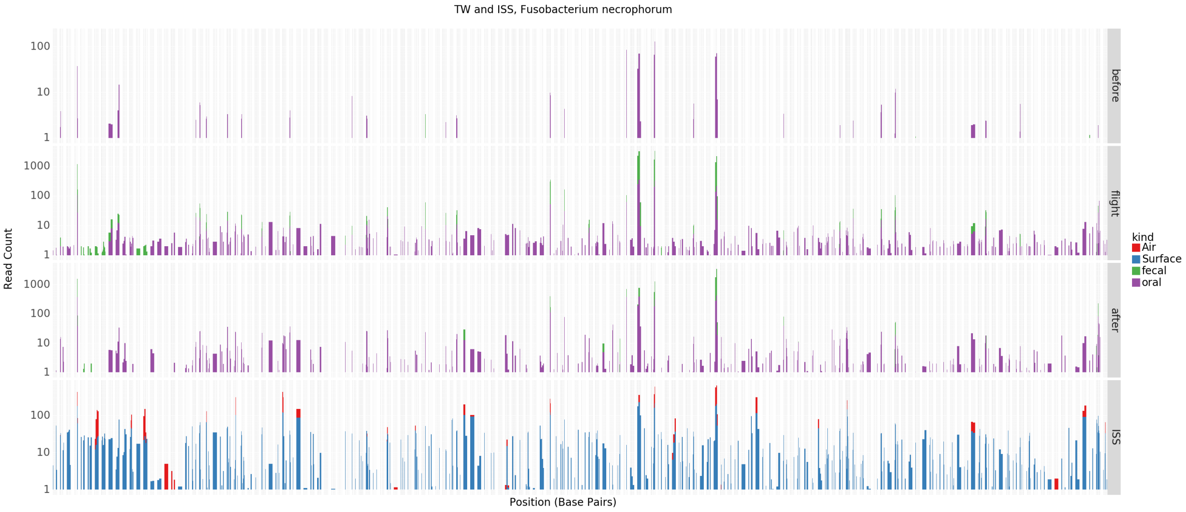

Figure S5: Rows show consolidated samples from before, during and after flight (or from the ISS at any point) from TW. Columns represent all available contigs for taxon. Colored bars represent 100bp covered, on average, at the specified read depth. A number of contigs are only covered in TW during and after flight.