Investigating the longitudinal dynamics of the human gut microbiome and immune system

by

Haixin Sarah Bi

B.Sc., University of Toronto (2016)

Submitted to the Graduate Program in Computational and Systems Biology
In partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Computational and Systems Biology

at the

Massachusetts Institute of Technology

February 2024

Signature of Author…….
Computational and Systems Biology Graduate Program
January 15, 2024

Certified by……...
Eric J. Alm
Professor of Biological Engineering

Accepted by…….
Christopher Burge
Professor of Biology
Co-Director, Computational and Systems Biology Graduate Program

Investigating the longitudinal dynamics of the human gut microbiome and immune system

by

Haixin Sarah Bi

Submitted to the Computational and Systems Biology Program
on January 15, 2024 in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Computational and Systems Biology

ABSTRACT

This thesis reports three explorations of two human-associated biological systems: the gut microbiome and the adaptive immune system. These two systems are closely intertwined, and engage in significant crosstalk with each other and with the rest of the body. Advances in recent years in high-throughput sequencing technologies have enabled study of these systems with unprecedented depth and comprehensiveness. In this thesis I leverage these tools to uncover novel insights into their composition and function. In the first chapter, I describe a project in which we closely tracked a small cohort of subjects' microbiomes over a period of up to one year, as they traveled from North America to Africa and back. We then intersected these data with data from African locals and contextualized our results via reanalysis of larger, published datasets on travelers' microbiomes, to build a fuller picture of what happens to our gut microbes when we travel. In the second chapter, I describe an analysis of a forthcoming microbiome data resource from our lab, in which I examined and quantified the diversity of antiphage defense capabilities at the strain level within the gut microbiome. My results underscore the value of large, cross-sectional datasets to capture underlying strain heterogeneity. Finally, in the third chapter, I describe a longitudinal study of an important component of adaptive immunity, the T cell receptor (TCR) repertoire, in healthy individuals over a time span of up to one year. Our results served as the first reference point for normal temporal fluctuations in TCR repertoire dynamics. Together, this thesis underscores the value of modern sequencing-based observational study of these complex human systems, using both cross-sectional and longitudinal designs.

Thesis supervisor: Eric J. Alm
Title: Professor, Biological Engineering

Acknowledgements

I gratefully acknowledge the following people and dog for their support, guidance, and good company during my graduate school tenure:

My advisor Prof. Eric Alm
My committee members, Profs. Doug Lauffenburger, Alex Shalek, and Fuqing Wu
Prof. Chris Burge and Jacquie Carota of the CSB program
Alm lab members past and present
CSB classmates
Dr. Rebecca Kusko
Friends and roommates in Cambridge and in Canada
Dr. Adam Ball
My parents, Jing and Xun
and finally, Charlie the dog

Introduction

The human microbiome comprises a diverse and dynamic community of microorganisms living in and on our bodies. The gastrointestinal tract, in particular, is home to the highest density of these microbes, numbering by some estimates in the tens of trillions (making it comparable to the number of cells in the human body), and being made up of up to several hundred distinct species (Ursell et al. 2012; Costea et al. 2018). Much interest and study in recent years have gone into several fundamental questions about these gut microbes. Who are they? What do they do? Where do they come from? When do they show up or go away? From these efforts, we have an idea of what a typical gut microbiome profile might look like (especially in better-studied western populations, although sampling inclusivity has been improving in recent years), what these microbes use up and output (and in turn, how these things affect and are affected by their human host bodies), and how mutable factors such as age, disease, or diet might be reflected in the gut (Rinninella et al. 2019). Advancements in next-generation sequencing technologies have facilitated the study of gut microbes at the population level. Common approaches to this will often begin with taking a stool sample, then purifying out DNA, and sequencing either everything (shotgun metagenomic sequencing) or a targeted marker gene that is conserved across all bacteria (the 16S ribosomal DNA locus is a popular choice). Improved whole genome sequencing platforms and protocols have also paved the way toward both a higher number and quality of reference genomes for diverse gut bacterial strains derived from single-isolate culturing techniques, popularizing along with it the concept of the "pangenome" of a species (Medini et al. 2005). As researchers continue to push the boundaries of microbiome science, more effort is going into scaling up these studies, in particular to draw more robust conclusions from observational designs.

Inextricably linked with the gut microbiome is the human immune system, a complex, highly-specialized group of bodily structures, including organs, vessels, cells, and macromolecules that together serve to detect and defend the body from harmful "non self" invaders (e.g. bacteria, viruses) and "altered self" components (e.g. damaged or

cancerous cells) (Murphy et al. 2008). At the same time, the immune system has mechanisms to communicate with and regulate symbionts such as commensals in the gut to allow coexistence. Human immunity comprises two branches, innate and adaptive, which together provide layered and extensive response mechanisms. The adaptive immune system has the distinction of being able to mount highly specific responses upon recognizing “non self” or “altered self” molecular signatures, termed antigens, and to retain memory of previous recognition events which affect the strength and timing of future repeat recognition events. This capacity is derived from the presence of specialized receptors on two types of adaptive immune cells, B and T cells, functioning in concert with cell surface proteins called Human Leukocyte Antigen (HLA) proteins. These B and T cell populations start out with already immensely diverse antigen recognition capacities from the get-go during human development. Subsequent antigen encounters throughout life further shape the collective antigen recognition capacities of these cells, termed the BCR or TCR repertoires. In recent years, interest in studying these adaptive immune repertoires has accelerated, as new sequencing protocols have been developed that can capture their diversity, sometimes tied to other rich data such as cell subtype identity and gene expression (Rosati et al. 2017; Pai and Satpathy 2021).

This thesis includes three chapters exploring the two systems described above. I will briefly introduce these chapters here.

Chapter 1 describes a project in which we tracked the gut microbiomes of four travelers from North America to Africa over the time span of up to a year, with daily or near daily sampling. Our project follows up on an earlier paper from our lab characterizing the two subjects’ gut (and oral) microbiomes over one year, one of whom happened to travel internationally during the sampling period (David et al. 2014). That paper, nearly a decade old, was a milestone in the burgeoning study of temporal variation in the human gut, and to our knowledge, a similar travel study design has not been presented since, likely due to the logistical challenge of frequent sampling, especially when subjects are abroad. In the project presented here, we additionally compared and contextualized our subjects’ time series with cross-sectional destination-matched cohorts of African locals

for whom we have microbiome data as well as a reanalysis of published travelers' microbiome datasets from the literature.

Chapter 2 describes an analysis of a forthcoming microbiome data resource from our lab, which I mined to identify antiphage defense systems and subsequently quantified their heterogeneity at the strain level, in order to demonstrate the value of our resource and to underline the hidden diversity in these systems likely still to be found.

Chapter 3 revisits the time series theme, and describes a project in which we conducted T cell receptor sequencing of blood samples from three subjects whom we followed longitudinally for one year. These data allowed us to study baseline diversity and temporal flux patterns the immune repertoires of healthy adults in a time span relevant for short-term changes such as seasonal infectious disease. We further characterize the "publicness" of observed TCRs by intersecting our data with a similar, large cross-sectional published dataset of TCR sequences.

References

Costea, Paul I., Falk Hildebrand, Manimozhiyan Arumugam, Fredrik Bäckhed, Martin J. Blaser, Frederic D. Bushman, Willem M. de Vos, et al. 2018. "Enterotypes in the Landscape of Gut Microbial Community Composition." *Nature Microbiology* 3 (1): 8–16.

David, Lawrence A., Arne C. Materna, Jonathan Friedman, Maria I. Campos-Baptista, Matthew C. Blackburn, Allison Perrotta, Susan E. Erdman, and Eric J. Alm. 2014. "Host Lifestyle Affects Human Microbiota on Daily Timescales." *Genome Biology* 15 (7): R89.

Medini, Duccio, Claudio Donati, Hervé Tettelin, Vega Masignani, and Rino Rappuoli. 2005. "The Microbial Pan-Genome." *Current Opinion in Genetics & Development* 15 (6): 589–94.

Murphy, Kenneth P., Kenneth M. Murphy, Paul Travers, Mark Walport, and Charles Janeway. 2008. *Janeway's Immunobiology*.

Pai, Joy A., and Ansuman T. Satpathy. 2021. "High-Throughput and Single-Cell T Cell Receptor Sequencing Technologies." *Nature Methods* 18 (8): 881–92.

Rinninella, Emanuele, Pauline Raoul, Marco Cintoni, Francesco Franceschi, Giacinto Abele Donato Miggiano, Antonio Gasbarrini, and Maria Cristina Mele. 2019. "What Is the Healthy Gut Microbiota Composition? A Changing Ecosystem across Age, Environment, Diet, and Diseases." *Microorganisms* 7 (1). https://doi.org/10.3390/microorganisms7010014.

Rosati, Elisa, C. Marie Dowds, Evaggelia Liaskou, Eva Kristine Klemsdal Henriksen,

Tom H. Karlsen, and Andre Franke. 2017. “Overview of Methodologies for T-Cell Receptor Repertoire Analysis.” *BMC Biotechnology* 17 (1): 61.
Ursell, Luke K., Jessica L. Metcalf, Laura Wegener Parfrey, and Rob Knight. 2012. “Defining the Human Microbiome.” *Nutrition Reviews* 70 Suppl 1 (Suppl 1): S38–44.

Chapter 1: Temporal dynamics of the human gut microbiome during international travel

This chapter is a collaboration between me, Mathilde Poyet, and Mathieu Groussin. Mathilde and Mathieu performed the study design and most of the library preparation for sequencing. I performed some of the library preparation with Amy Xiao and all of the analyses and visualization. This chapter is written by me.

Background

With the broad availability and increased affordability of high-throughput sequencing technologies, large-scale interrogations of human host-associated microbial communities have become widespread. Much of this interest is driven by our deepening understanding of the inextricable two-way link between microbiome and host health. In addition to large cross-sectional studies, research efforts in the past two decades have explored how human microbiomes change over time (Palmer et al. 2007; Caporaso et al. 2011; Faith et al. 2013; Rajilić-Stojanović et al. 2012). From these landmark longitudinal studies, we know that there generally tends to be individuality and stability that is maintained on the timescale of months to years to decades. Despite this, it has been shown that lifestyle factors such as alterations in diet, stress, and medication use can induce shifts in microbiome away from these stable states (Vich Vila et al. 2020; Singh et al. 2017; Madison and Kiecolt-Glaser 2019). It remains an ongoing area of research which of these lifestyle factors result in what changes in the microbiome, whether these changes are temporary or permanent, and whether these changes are beneficial or otherwise for the host.

One such important lifestyle factor is host travel. Each year, over one billion international tourist arrivals are logged globally, and hundreds of millions of people travel from industrialized to resource-limited, tropical or semitropical countries (Kampmann et al. 2021). The frequency of modern travel makes it relevant for understanding both microbiome and host health outcomes on an individual scale and also public health at large. On an individual scale, previous studies have demonstrated travel-associated effects such as broad microbial dysbiosis as well as upticks in diarrheagenic and multi-drug resistant taxa (David et al. 2014; Youmans et al. 2015; Kennedy and Collignon 2010). The latter presents a burgeoning challenge to global health as travel is known to facilitate the exchange of antimicrobial resistance (AMR) between high- and low-prevalence populations (D’Souza et al. 2021).

Untangling the effects of travel on the gut microbiome is a complex task, since travel usually comprises many concomitant changes – to diet, to activity levels, to stress, to

circadian rhythms, and to one's environment. Few study designs can address all of these factors.

Existing work on the impact of international travel on gut microbiome dynamics have largely been limited to pre- and post-travel or otherwise sparse sampling designs (Cheung et al. 2023; Youmans et al. 2015; D'Souza et al. 2021; Langelier et al. 2019; Bengtsson-Palme et al. 2015; Leo et al. 2019; Pires et al. 2019; Kampmann et al. 2021; Peng et al. 2021). Given the difficulty of during-travel sample acquisition, to our knowledge the only prior study to include dense intra-subject sampling comprising pre-, during-, and post-travel time periods is from our lab (David et al. 2014). Our early findings then demonstrated that microbiome changes happen on a timescale of days. Furthermore, many of the studies on travelers' gut microbiota only examine certain taxa of interest (usually antimicrobial-resistant Enterobacteriaceae) and thus use culture- or targeted PCR-based approaches (Mellon et al. 2020; Paltansing et al. 2013; Blyth et al. 2016; Tängdén et al. 2010; Ostholm-Balkhed et al. 2013; Kennedy and Collignon 2010; von Wintersdorff et al. 2016; Kantele et al. 2015; Reuland et al. 2016). A few recent studies that do employ high-throughput community sequencing-based approaches also take a narrow focus on resistome changes (D'Souza et al. 2021; Boolchandani et al. 2022). Little attention has been given to broader gut community dynamics of traveling hosts beyond immediate disease and AMR associations.

Our aim with this study is to follow up on the earlier microbiome time series work from our lab and to expand its scope. We follow a modestly bigger cohort of subjects for whom we analyze long, densely-sampled microbiome time series. Subjects all originated in an urban, western geography and variously traveled to different countries in Africa, visiting both towns and cities as well as rural locales. Multiple subjects share travel destinations in Africa and for several destinations we additionally have matching large local subject cohorts whose microbiomes we also profiled via community sequencing. We validate some of our results from our new, densely-sampled but limited sized cohort data via analysis of previously-published large cohort sequencing studies of travelers' microbiomes (with different data types) to provide independent lines of evidence.

Methods

Four healthy adult subjects were included for participation, all of whom for the duration of the study resided in Massachusetts, USA and had planned travel to various African countries (**Figure 1a**). Subjects were instructed on a wipe-based self-collection protocol for stool samples as previously described (García 2018), and collected samples daily or as frequently as possible starting prior to travel, continuing through travel periods, and following their return to the United States (**Figure 1b**). Subjects self-recorded their location and illness status on each day of sample collection. Subject time series ranged from 45 to 251 days long.

Stool samples were flash-frozen when collected during travel or otherwise stored within hours at -80C when collected in Boston. Sample processing and DNA extraction were performed using the Qiagen PowerSoil kit as previously described, and paired-end Illumina libraries were prepared for 16S amplicon sequencing. Sequencing was performed at the Broad Institute Microbial ‘Omics Core (for batches 1 and 2) and the BioMicro Center at MIT (for batches 3 and 4). Samples were assigned to batches chronologically.

Raw sequencing data were processed using Qiime2 software v. 2021.4 (Bolyen et al. 2019). Amplicon sequence variants (ASVs) were denoised with Dada2 and taxonomic labels were assigned using the naive Bayes classifier pre-trained on the Silva 138 99% OTUs from 515F/806R 16S region provided by Qiime2 developers (Bokulich et al. 2018; Robeson et al. 2021). Subsequent analysis and visualization were performed with custom Python and R scripts.

A large cross-sectional dataset of microbiome 16S amplicon sequencing profiles from stool samples from locals in some of the African destination countries as well as the US was previously curated in the lab. A previously-generated feature table and associated taxonomy table collapsed at the genus level were used for this study.

Published datasets were identified via a literature search and downloaded from the Sequence Read Archive. Shotgun metagenomic sequencing data (Cheung et al.,

D'Souza et al., and Boolchandani et al.) were processed using Kraken2 and Bracken software (Wood, Lu, and Langmead 2019; Lu et al. 2017). A standard Kraken2 database comprising Refseq archaeal, bacterial, viral, plasmid, and human reference sequences and UniVec vector sequences dating from Mar 2023 available on benlangmead.github.io was used, as well as an accompanying Bracken database from the same update. 16S amplicon data (Pires et al. and David et al.) were processed using Qiime2, same as above. Analysis and visualization were performed in Python v. 3.9.7.

Results

Microbiome diversity declined following travel periods across diverse destinations

We first examined broad microbial community diversity throughout subjects' time series. A handful of previous studies have found no statistically-significant difference in intra-individual diversity before and after travel, however these conclusions came from small sample sizes (as small as n=10), so there is yet to be clarity on whether travel is robustly associated with any directional shift in diversity as larger studies have only recently come out, nor has there been a cross-study comparison (Youmans et al. 2015; Leo et al. 2019; Langelier et al. 2019). We thus explored this question first in our novel data, and then in recently published large(r) cohort study datasets. In our study we observed that microbiome alpha diversity as measured by Shannon index modestly decreased in ¾ travelers after returning from Africa (mean pre-travel=6.08, mean post-travel=5.99), averaging across daily samples during each period (**Figure 2a**). Diversity trajectories were inconsistent among travelers during travel, again averaging across samples during this period. Comparing with our previous work (David et al. 2014), we see the same slight decline in alpha diversity in the traveling subject (Subject A) during the sampling period (mean pre-travel=4.38, mean post-travel=4.33) (**Figure 2b**).

We observed similar and more robust findings in the large cohort published datasets included in our analysis. We analyzed three studies (Cheung et al., D'Souza et al., and

Pires et al.) with pre- vs. post-travel sampling designs (the former two being the two largest cohort travel studies with public data of this design we identified, among over a dozen studies). Among the Cheung et al. travelers (n=90, originating in Hong Kong and visiting various destinations mostly in Southeast and East Asia), there was a significant overall downturn in diversity post-return (Wilcoxon signed-rank test p=0.033, 55% of travelers experienced absolute declines in diversity) (**Figure 3a**). We observed the same outcome in the D'Souza et al. Dutch travelers (n=190, Wilcoxon signed-rank test p=0.007, 58% of travelers declined) (**Figure 3b**). Since subject-specific destination information was available for these data, we grouped subjects by regional destination and repeated testing for each group, however with these smaller cohort sizes none were standalone significant (Southeastern Asia p=0.17; Southern Asia p=0.24; Northern Africa p=0.20; Eastern Africa p=0.12) (**Figure 3e**). Next we analyzed the Pires et al. Swiss traveler data (n=40, all subjects travel to India), but comparing the pre-travel and immediately-post travel timepoints, we did not observe a significant diversity shift, possibly due to the smaller cohort size (Wilcoxon signed-rank test p=0.38, comparing subject-matched pre- and immediately-post travel samples). Anecdotally, however, there was a minor and gradual long term uptick in diversity (mean pre-travel=7.04, mean immediately post-travel=7.18, mean 3 months post-travel=7.47, mean 6 months post-travel=7.56, mean 12 months post-travel=7.31) (**Figure 3c**).

To dive deeper into in-travel diversity dynamics, we analyzed the Boolchandani et al. data (n=159, comprising travelers with many countries of origin, all traveling to Peru and sampled exclusively during their time in Peru), we did not observe a clear directional change in diversity during the course of travel (fitted linear regression model shannon_index = 0.002*day_count + 5.84, R^2=0.004), however this study did not include pre- or post-trip samples for us to compare with other datasets (**Figure 3d**). Thus our analyses showed a recurring pattern of microbiome diversity decrease immediately post-travel and that this held across diverse geographies for traveler origins and destinations, but that the timing of the onset of this decrease was inconsistent (although there is not yet sufficient data to address this).

High-level microbiome structure was maintained but minor taxa experienced temporary blooms

We next explored community compositional changes brought on by travel in our subjects' time series. All four of our travelers' microbiomes were dominated by Firmicutes and secondarily Bacteroidetes, and this hierarchy held stable through the sampling period (**Figure 4**). In contrast with our earlier findings (David et al. 2014) but consistent with other published work (Youmans et al. 2015), Firmicutes-to-Bacteroidetes ratio temporarily increased during African travel periods and then reversed in the majority (¾) of our subjects (**Figure 5**). Subject 1, who was the exception, was also the only subject to visit Africa twice during the sampling period, and both times experienced a decrease in Firmicutes-to-Bacteroides ratio (comparing with the adjacent North American time periods for each African trip), which remained lower post-travel.

In two of our subjects, we observed blooms in already-present Verrucomicrobiota coinciding with African travel (**Figure 4**), which, in subject 1, who has the longer time series, does not persist or reoccur beyond the end of their first period in Africa. We observed as well that average relative abundance of Proteobacteria decreased in ¾ of subjects going from pre- to during-travel, but subsequently increased in all subjects, resulting in overall higher abundance post- versus pre-travel in half of subjects (**Figure 6**). We saw consistent outcomes in our analysis of published data. Among the Pires et al. travelers, major gut phyla remained steady while minor phyla (Proteobacteria, Verrucomicrobiota, and Fusobacteriota) experienced upswings immediately post-return, which were restored to pre-travel levels by the next sampling point and maintained out to 12 months post-travel (**Figure 7**). We were particularly interested in Proteobacteria dynamics in traveler microbiomes, as this phylum encompasses common illness-causing and antibiotic resistance-associated Enterobacteriaceae. Indeed, we observed spikes in Enterobacteriaceae in all four of our subjects which appeared to coincide with travel to Africa (**Figure 8**). In two of the three published datasets we used which included pre- and post-travel sampling time points, the two outcomes of lower post-travel microbiome diversity and higher abundance of Proteobacteria tended to coincide, though this was statistically significant only in the biggest study cohort

(D'Souza et al., Fisher's exact test $p=3.69*10^{-10}$) (**Figure 9**). This was also true on a smaller scale in our novel data. Together, these findings underscore the ubiquity of Proteobacteria abundance as a signature of post-travel microbiome disturbance.

We next dug into ASV-level dynamics within our traveler data. In concordance with results from our earlier study, we found that mean relative abundance of ASVs was positively correlated with their temporal persistence across all subjects, and that there was a core set of abundant and prevalent taxa for each person (**Figure 10**). We examined the dynamics of the most abundant and prevalent ASVs in each subject's time series and observed that many experienced temporary downturns during travel periods, coincident with when minor taxa bloomed (**Figure 11**). We recapitulated the analysis on conditionally-rare taxa (CRT) from our earlier work using the same thresholds (Gibbons et al. 2017) but identified very few CRTs in any time series (ranging in count from 0-11 among subjects), however interestingly despite this, three subjects had CRTs in common (**Figure 12a**). CRT blooms appear to coincide with travel periods in some but not every case (**Figure 12b**).

Broad microbiome shifts occurred during travel but were largely independent of specific traveler destination

We next examined overall microbiome shifts through time among our travelers. Above all else, we observed that our travelers maintained individuality in their microbiome signatures over time, as samples clustered most clearly by subject, compared to geography or processing batch (**Figure 13**). We hypothesized that there may have been broad shifts concomitant with travel, and indeed we observed an "out-and-back" pattern in subjects 1, 2, and 4, wherein their in-Africa samples appeared more disparate compared to their in-North America and in-Europe samples, both before and after African travel (**Figure 14**). In the Pires et al. travelers, we anecdotally observed a slight long-term divergence from pre-travel sampling points, however we cannot directly compare patterns in trajectory without in-travel samples (**Figure 15**).

With our destination-matched local cohort data, we wanted to examine whether our travelers' microbiomes became more similar to those of local residents from the African

countries they visited while they were there. Specifically, our local cohort included samples from Cameroon (uniquely-sampled donors n=103, visited by subjects 1 and 2), Tanzania (n=91, visited by all subjects), and Rwanda (n=92, visited by subject 1), as well as from the United States (n=117). We did not observe a clear “assimilation” pattern for any of the three African countries, however for subjects 1, 2, and 4 there appeared to be shifts away from American local microbiome profiles, such that these travelers’ gut communities became more dissimilar versus their country of origin while simultaneously not becoming more similar to those of their temporary host countries (**Figure 16**).

We further explored this idea of destination-specific effects using our collection of published datasets. With the D’Souza et al. data we asked whether there was more or less divergence from pre-travel baseline depending on destination geography, however we did not observe a significant difference between the four regions (Kruskal-Wallis test p=0.74) (**Figure 17**). We also wanted to examine whether same-destination post-travel microbiomes of distinct subjects became more similar to each other, compared to different-destination subjects. In other words, was there a “convergence” by destination? We did not observe this in the D’Souza et al. cohort, comparing all-versus-all post-travel samples for destination-matched and mismatched subjects (**Figure 18**). Approaching the question of “convergence” from a different angle, we sought to determine whether travelers with shared destination were more similar to each other after travel than they were before (perhaps a “shared experience effect”, which in the Pires et al. data appeared anecdotally to be the case, interestingly with continuing convergence further out from return date (**Figure 19**). Taken together, the bulk of our results suggest that travel-associated microbiome shifts are largely decoupled from specific traveler destination.

There was no enrichment in local microbe colonization in Africa, and in-Africa colonization events were not distinctly different from others elsewhere

Finally we focused on the question of new microbe influx during travel. Our expectation was that we might see distinct local signatures of microbe colonization while in Africa, however we did not observe this in our results. For each of our subjects, we identified every ASV first observed while in Africa, which presumably were taken up from

subjects' environs. Firstly, we observed little overlap between subjects of these newly-acquired ASVs, despite shared destinations and travel experiences among some subjects (indeed, we observed more overlap among ASVs lost in Africa) (**Figure 20a-b**). The phylum distribution of newly-acquired ASVs reflected dominant phyla in subjects' existing microbiome composition (**Figure 20c)**, and did not include any largely-exclusive-to-Africa microbes from the cross-sectional dataset (**Figure 21**). With our destination-matched cross-sectional sampling cohort, we tested whether these new microbes were shared with local African populations, which one might expect to see given proximity and sharing of diets. Unexpectedly, we did not see this effect (at the genus level), and there was no marked enrichment for shared microbes among all new microbes first appearing in subject time series in Africa, comparing between destination country cohorts (Cameroon, Tanzania, and Rwanda) and unrelated country cohorts (**Figure 22**). Of note was the fact that dominant phyla overlapped between our travelers and our local African cohorts, and where they differed were mostly in minor phyla.

We next examined the frequency of colonization through travel periods. Segmenting each time series by geographic region, we calculated the average number of new ASVs per day per continuous time period in each region. In the results, we did not see consistently outsized colonization frequency during periods in Africa for our subjects (**Figure 23**). Lastly, we asked whether, for significantly-long continuous time periods in each geography for each subject, there was any difference in temporal persistence or peak or mean abundance of newly-colonized microbes. In plain terms, we wanted to ask whether there was anything special about microbes acquired in non-native surroundings that allowed them to colonize particularly well (or poorly). We did not observe a clear association across subjects between geography and the proportion of new microbes being less (≤3 sampling days) or more (>3 sampling days) prevalent (**Figure 24**), nor between geography and the two abundance measures (**Figures 25-6**). Examining the temporal dynamics of the cumulative abundance of Africa-acquired microbes, we observed, across all subjects, a temporary spike immediately post-colonization followed by a drop to a sustained low abundance post-travel (**Figure 27**). Only in Subject 2 did Africa-acquired microbes appear to establish a meaningful (but still minor) foothold.

Discussion

In this study we investigated the impact of international travel from more developed to less developed geographies on the human gut microbiome. By complementing our analysis of our novel long and dense time series with reanalysis of large, cross-sectional studies, we uncovered new insights and delivered them with broader context. We found that diminished microbiome diversity was significantly associated with return from travel, and that this effect was observed across diverse destinations in developing regions. This runs counter to earlier negative conclusions from more limited investigations (Kampmann et al. 2021; Langelier et al. 2019; Leo et al. 2019), and suggests a new lower-bound to the time it takes to observe an effect on microbiota diversity post-relocation. It is possible that traveling between highly dissimilar environments may prompt a fast adjustment or perturbation, perhaps coinciding with a shift in diet or behaviour. As a point of comparison, earlier work studying immigrants to the US from Southeast Asia reported that diversity measures dropped within 9 months of permanently moving, presumably reflecting a longer-term acclimatization (Vangay et al. 2018).

In terms of microbiome composition, we observed mostly temporary effects from travel in the form of fluctuations of lesser-abundant phyla. Our findings corroborated earlier reports of temporarily-elevated Enterobacteriaceae resulting from travel to developing regions and confirmed that it occurs often and irrespective of specific geography (Kampmann et al. 2021; Leo et al. 2019; Langelier et al. 2019). This suggested that there were common triggers stemming from travel-associated lifestyle change and exposure to non-native surroundings.

The microbiome shifts we saw in our analyses appeared to be largely temporary, and did not differ meaningfully by destination. Nor did, surprisingly, within-travel microbial colonization patterns. Indeed, while there were in some subjects significant numbers of newfound microbes in Africa, they did not seem to originate from substantial exchange with local gut microbiome reservoirs. Instead, we saw a “more of the same” outcome emerge. It would be difficult to pinpoint exactly why this was the case, but we speculate

that perhaps it would take a more extended duration stay for subjects to meaningfully shift toward a more “local” state (including changes such as adopting a more local diet), like what was observed in the immigration study mentioned earlier (Vangay et al. 2018). It could also be that microbes we labeled as “new” during this period were simply already present but below the limit of detection before then.

We acknowledge some limitations of our study, chiefly the limited cohort size of our primary dataset. Our subjects were in proximity to each other for the duration of sampling, and in some instances, traveled together as a party. Our subjects were also sometimes in proximity to subjects in the local cross-sectional cohort whose data we intersected with our longitudinal data. More generally, in each case we aimed to recapitulate our findings in large published datasets, however much of this was not possible as there are no other directly comparable studies besides the smaller one from our group years ago (David et al. 2014). Thus, we cannot speculate on the extent to which our observations from our cohort of 4 will generalize more broadly. This study represents a preliminary step towards this kind of heavy repeated-sampling-based interrogation of microbiome dynamics in travel, which can address questions that simple pre- versus post- sampling designs cannot. Further scaling up is a formidable challenge, and would likely require non-trivial subject education and compliance on self-sampling procedures.

Altogether, we hope that our findings here will fill in some gaps in the current understanding of microbiome dynamics in travel, and pave the way to more comprehensive and rigorously-designed studies in the future.

References

Bengtsson-Palme, Johan, Martin Angelin, Mikael Huss, Sanela Kjellqvist, Erik Kristiansson, Helena Palmgren, D. G. Joakim Larsson, and Anders Johansson. 2015. “The Human Gut Microbiome as a Transporter of Antibiotic Resistance Genes between Continents.” *Antimicrobial Agents and Chemotherapy* 59 (10): 6551–60.

Blyth, Dana M., Katrin Mende, Ashley M. Maranich, Miriam L. Beckius, Kristie A. Harnisch, Crystal A. Rosemann, Wendy C. Zera, Clinton K. Murray, and Kevin S. Akers. 2016. “Antimicrobial Resistance Acquisition after International Travel in U.S.

Travelers.” *Tropical Diseases, Travel Medicine and Vaccines* 2 (March): 4.
Bokulich, Nicholas A., Benjamin D. Kaehler, Jai Ram Rideout, Matthew Dillon, Evan Bolyen, Rob Knight, Gavin A. Huttley, and J. Gregory Caporaso. 2018. “Optimizing Taxonomic Classification of Marker-Gene Amplicon Sequences with QIIME 2’s q2-Feature-Classifier Plugin.” *Microbiome* 6 (1): 90.
Bolyen, Evan, Jai Ram Rideout, Matthew R. Dillon, Nicholas A. Bokulich, Christian C. Abnet, Gabriel A. Al-Ghalith, Harriet Alexander, et al. 2019. “Reproducible, Interactive, Scalable and Extensible Microbiome Data Science Using QIIME 2.” *Nature Biotechnology* 37 (8): 852–57.
Boolchandani, Manish, Kevin S. Blake, Drake H. Tilley, Miguel M. Cabada, Drew J. Schwartz, Sanket Patel, Maria Luisa Morales, et al. 2022. “Impact of International Travel and Diarrhea on Gut Microbiome and Resistome Dynamics.” *Nature Communications* 13 (1): 7485.
Caporaso, J. Gregory, Christian L. Lauber, Elizabeth K. Costello, Donna Berg-Lyons, Antonio Gonzalez, Jesse Stombaugh, Dan Knights, et al. 2011. “Moving Pictures of the Human Microbiome.” *Genome Biology* 12 (5): R50.
Cheung, Man Kit, Rita W. Y. Ng, Christopher K. C. Lai, Chendi Zhu, Eva T. K. Au, Jennifer Wing Ki Yau, Carmen Li, et al. 2023. “Alterations in Faecal Microbiome and Resistome in Chinese International Travellers: A Metagenomic Analysis.” *Journal of Travel Medicine*, March. https://doi.org/10.1093/jtm/taad027.
David, Lawrence A., Arne C. Materna, Jonathan Friedman, Maria I. Campos-Baptista, Matthew C. Blackburn, Allison Perrotta, Susan E. Erdman, and Eric J. Alm. 2014. “Host Lifestyle Affects Human Microbiota on Daily Timescales.” *Genome Biology* 15 (7): R89.
D’Souza, Alaric W., Manish Boolchandani, Sanket Patel, Gianluca Galazzo, Jarne M. van Hattem, Maris S. Arcilla, Damian C. Melles, et al. 2021. “Destination Shapes Antibiotic Resistance Gene Acquisitions, Abundance Increases, and Diversity Changes in Dutch Travelers.” *Genome Medicine* 13 (1): 79.
Faith, Jeremiah J., Janaki L. Guruge, Mark Charbonneau, Sathish Subramanian, Henning Seedorf, Andrew L. Goodman, Jose C. Clemente, et al. 2013. “The Long-Term Stability of the Human Gut Microbiota.” *Science* 341 (6141): 1237439.
García, Mariana Guadalupe Matus. 2018. *Analysis of Fecal Biomarkers to Impact Clinical Care and Public Health*.
Gibbons, Sean M., Sean M. Kearney, Chris S. Smillie, and Eric J. Alm. 2017. “Two Dynamic Regimes in the Human Gut Microbiome.” *PLoS Computational Biology* 13 (2): e1005364.
Kampmann, Christian, Johan Dicksved, Lars Engstrand, and Hilpi Rautelin. 2021. “Changes to Human Faecal Microbiota after International Travel.” *Travel Medicine and Infectious Disease* 44 (November): 102199.
Kantele, Anu, Tinja Lääveri, Sointu Mero, Katri Vilkman, Sari H. Pakkanen, Jukka Ollgren, Jenni Antikainen, and Juha Kirveskari. 2015. “Antimicrobials Increase Travelers’ Risk of Colonization by Extended-Spectrum Betalactamase-Producing Enterobacteriaceae.” *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America* 60 (6): 837–46.
Kennedy, K., and P. Collignon. 2010. “Colonisation with Escherichia Coli Resistant to ‘Critically Important’ Antibiotics: A High Risk for International Travellers.” *European*

Journal of Clinical Microbiology & Infectious Diseases: Official Publication of the European Society of Clinical Microbiology 29 (12): 1501–6.
Langelier, Charles, Michael Graves, Katrina Kalantar, Saharai Caldera, Robert Durrant, Mark Fisher, Richard Backman, Windy Tanner, Joseph L. DeRisi, and Daniel T. Leung. 2019. “Microbiome and Antimicrobial Resistance Gene Dynamics in International Travelers.” *Emerging Infectious Diseases* 25 (7): 1380–83.
Leo, Stefano, Vladimir Lazarevic, Nadia Gaïa, Candice Estellat, Myriam Girard, Sophie Matheron, Laurence Armand-Lefèvre, Antoine Andremont The VOYAG-R study group, Jacques Schrenzel, and Etienne Ruppé. 2019. “The Intestinal Microbiota Predisposes to Traveler’s Diarrhea and to the Carriage of Multidrug-Resistant Enterobacteriaceae after Traveling to Tropical Regions.” *Gut Microbes* 10 (5): 631–41.
Lu, Jennifer, Florian P. Breitwieser, Peter Thielen, and Steven L. Salzberg. 2017. “Bracken: Estimating Species Abundance in Metagenomics Data.” *PeerJ. Computer Science* 3 (e104): e104.
Madison, Annelise, and Janice K. Kiecolt-Glaser. 2019. “Stress, Depression, Diet, and the Gut Microbiota: Human-Bacteria Interactions at the Core of Psychoneuroimmunology and Nutrition.” *Current Opinion in Behavioral Sciences* 28 (August): 105–10.
Mellon, Guillaume, Sarah E. Turbett, Colin Worby, Elizabeth Oliver, A. Taylor Walker, Maroya Walters, Paul Kelly, et al. 2020. “Acquisition of Antibiotic-Resistant Bacteria by U.S. International Travelers.” *The New England Journal of Medicine* 382 (14): 1372–74.
Ostholm-Balkhed, Ase, Maria Tärnberg, Maud Nilsson, Lennart E. Nilsson, Håkan Hanberger, Anita Hällgren, and Travel Study Group of Southeast Sweden. 2013. “Travel-Associated Faecal Colonization with ESBL-Producing Enterobacteriaceae: Incidence and Risk Factors.” *The Journal of Antimicrobial Chemotherapy* 68 (9): 2144–53.
Palmer, Chana, Elisabeth M. Bik, Daniel B. DiGiulio, David A. Relman, and Patrick O. Brown. 2007. “Development of the Human Infant Intestinal Microbiota.” *PLoS Biology* 5 (7): e177.
Paltansing, Sunita, Jessica A. Vlot, Margriet E. M. Kraakman, Romy Mesman, Marguerite L. Bruijning, Alexandra T. Bernards, Leo G. Visser, and Karin Ellen Veldkamp. 2013. “Extended-Spectrum β-Lactamase-Producing Enterobacteriaceae among Travelers from the Netherlands.” *Emerging Infectious Diseases* 19 (8): 1206–13.
Peng, Ye, Dengwei Zhang, Ting Chen, Yankai Xia, Peng Wu, Wai-Kay Seto, Anita L. Kozyrskyj, Benjamin J. Cowling, Jincun Zhao, and Hein M. Tun. 2021. “Gut Microbiome and Resistome Changes during the First Wave of the COVID-19 Pandemic in Comparison with Pre-Pandemic Travel-Related Changes.” *Journal of Travel Medicine* 28 (7). https://doi.org/10.1093/jtm/taab067.
Pires, João, Julia G. Kraemer, Esther Kuenzli, Sara Kasraian, Regula Tinguely, Christoph Hatz, Andrea Endimiani, and Markus Hilty. 2019. “Gut Microbiota Dynamics in Travelers Returning from India Colonized with Extended-Spectrum Cephalosporin-Resistant Enterobacteriaceae: A Longitudinal Study.” *Travel Medicine and Infectious Disease* 27: 72–80.

Rajilić-Stojanović, Mirjana, Hans G. H. J. Heilig, Sebastian Tims, Erwin G. Zoetendal, and Willem M. de Vos. 2012. “Long-Term Monitoring of the Human Intestinal Microbiota Composition.” *Environmental Microbiology*, October. https://doi.org/10.1111/1462-2920.12023.

Reuland, E. A., G. J. B. Sonder, I. Stolte, N. Al Naiemi, A. Koek, G. B. Linde, T. J. W. van de Laar, C. M. J. E. Vandenbroucke-Grauls, and A. P. van Dam. 2016. “Travel to Asia and Traveller’s Diarrhoea with Antibiotic Treatment Are Independent Risk Factors for Acquiring Ciprofloxacin-Resistant and Extended Spectrum β-Lactamase-Producing Enterobacteriaceae-a Prospective Cohort Study.” *Clinical Microbiology and Infection: The Official Publication of the European Society of Clinical Microbiology and Infectious Diseases* 22 (8): 731.e1–7.

Robeson, Michael S., 2nd, Devon R. O’Rourke, Benjamin D. Kaehler, Michal Ziemski, Matthew R. Dillon, Jeffrey T. Foster, and Nicholas A. Bokulich. 2021. “RESCRIPt: Reproducible Sequence Taxonomy Reference Database Management.” *PLoS Computational Biology* 17 (11): e1009581.

Singh, Rasnik K., Hsin-Wen Chang, Di Yan, Kristina M. Lee, Derya Ucmak, Kirsten Wong, Michael Abrouk, et al. 2017. “Influence of Diet on the Gut Microbiome and Implications for Human Health.” *Journal of Translational Medicine* 15 (1): 73.

Tängdén, Thomas, Otto Cars, Asa Melhus, and Elisabeth Löwdin. 2010. “Foreign Travel Is a Major Risk Factor for Colonization with Escherichia Coli Producing CTX-M-Type Extended-Spectrum Beta-Lactamases: A Prospective Study with Swedish Volunteers.” *Antimicrobial Agents and Chemotherapy* 54 (9): 3564–68.

Vangay, Pajau, Abigail J. Johnson, Tonya L. Ward, Gabriel A. Al-Ghalith, Robin R. Shields-Cutler, Benjamin M. Hillmann, Sarah K. Lucas, et al. 2018. “US Immigration Westernizes the Human Gut Microbiome.” *Cell* 175 (4): 962–72.e10.

Vich Vila, Arnau, Valerie Collij, Serena Sanna, Trishla Sinha, Floris Imhann, Arno R. Bourgonje, Zlatan Mujagic, et al. 2020. “Impact of Commonly Used Drugs on the Composition and Metabolic Function of the Gut Microbiota.” *Nature Communications* 11 (1): 362.

Wintersdorff, Christian J. H. von, Petra F. G. Wolffs, Julius M. van Niekerk, Erik Beuken, Lieke B. van Alphen, Ellen E. Stobberingh, Astrid M. L. Oude Lashof, Christian J. P. A. Hoebe, Paul H. M. Savelkoul, and John Penders. 2016. “Detection of the Plasmid-Mediated Colistin-Resistance Gene Mcr-1 in Faecal Metagenomes of Dutch Travellers.” *The Journal of Antimicrobial Chemotherapy* 71 (12): 3416–19.

Wood, Derrick E., Jennifer Lu, and Ben Langmead. 2019. “Improved Metagenomic Analysis with Kraken 2.” *Genome Biology* 20 (1): 257.

Youmans, Bonnie P., Nadim J. Ajami, Zhi-Dong Jiang, Frederick Campbell, W. Duncan Wadsworth, Joseph F. Petrosino, Herbert L. DuPont, and Sarah K. Highlander. 2015. “Characterization of the Human Gut Microbiome during Travelers’ Diarrhea.” *Gut Microbes* 6 (2): 110–19.

Figures

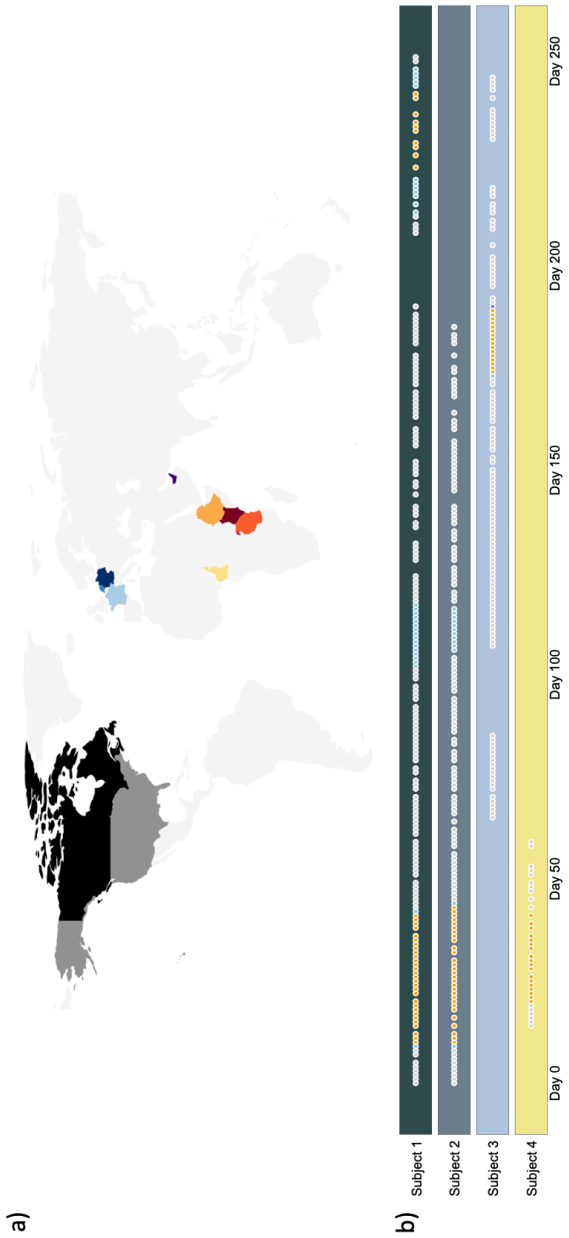


Figure 1. a) Map of subject origins and travel destinations. b) Subjects' self-sampling scheme. Each coloured circle denotes a sampling day, and circles are coloured by subjects' location (grey=North America, orange=Africa, Blue=Europe, Purple=Middle East).

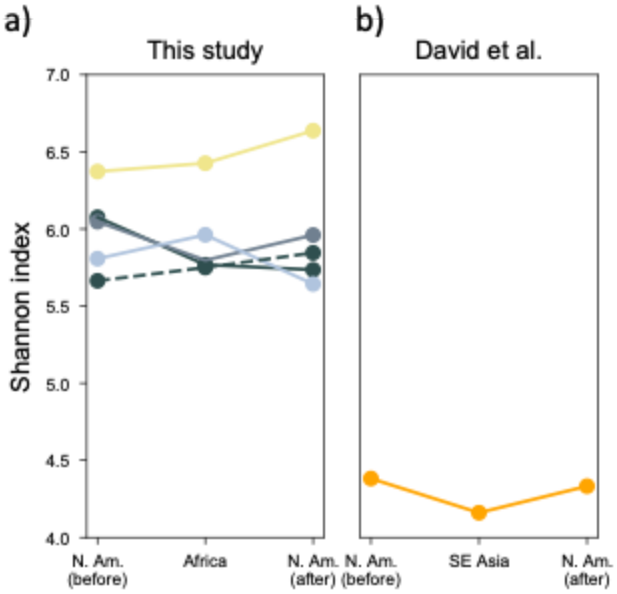


Figure 2. a) Alpha diversity trajectories (as measured by mean Shannon index in each continuous period in each region) across subject time series in this study. Dashed line denotes Subject 1's second trip to Africa, and their adjacent stays in North America. b) The same visualization for Subject A's time series from David et al., segmented by their time before, during, and after living abroad in Southeast Asia.

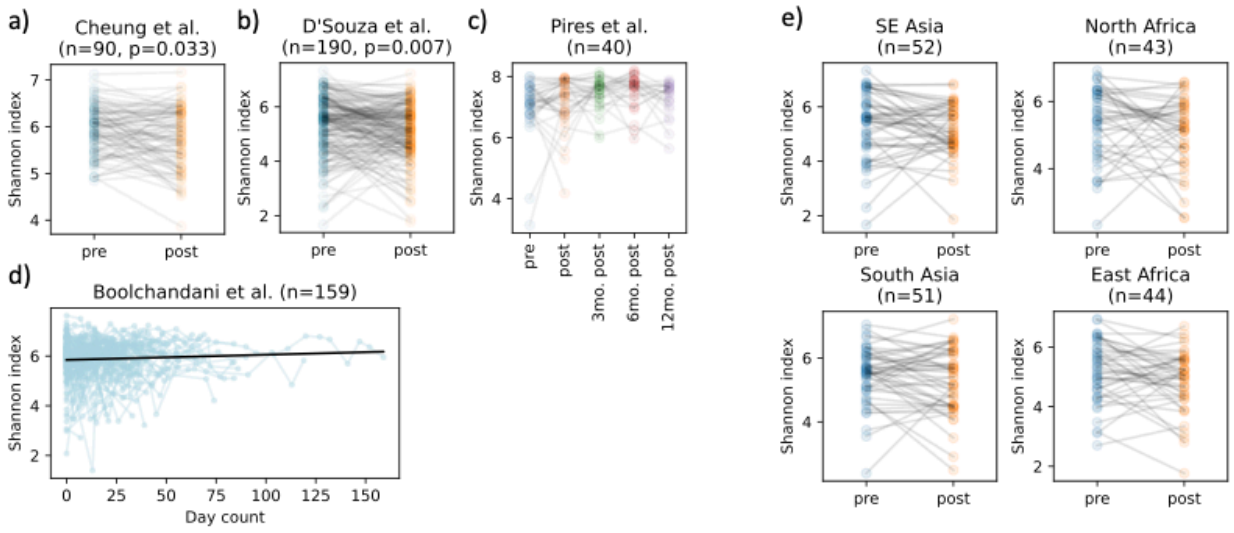


Figure 3. Subject-matched alpha diversity (as measured by Shannon index) before and after travel among the a) Cheung et al. and b) D'Souza et al. travelers. c) Subject-matched alpha diversity (Shannon index) at each of the 5 sampling points in Pires et al. d) Individual alpha diversity (Shannon index) trajectories of the Boolchandani

et al. travelers while in Cusco, Peru. e) same as b) but separating travelers by destination.

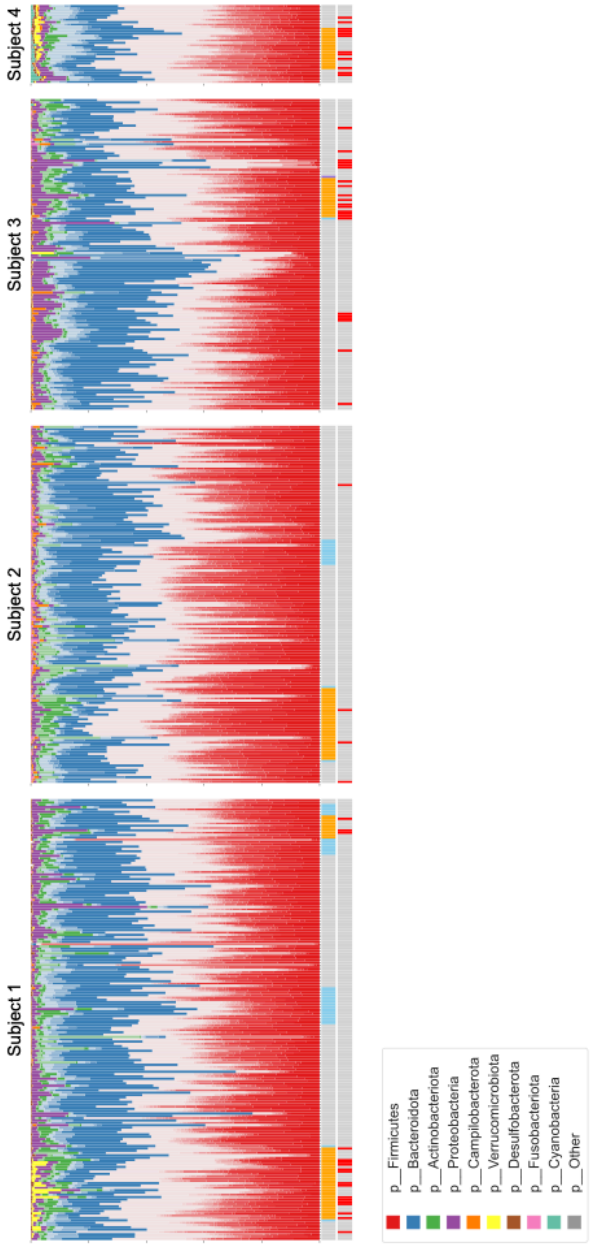


Figure 4. Stacked bar plots depicting subject microbiome composition over time. ASVs are grouped by genus, genuses belonging to a phylum are coloured as distinct shades of the colour assigned to that phylum (e.g. Firmicutes are shown in shades of red). Within a phylum, genuses are ordered bottom-to-top by descending relative abundance, and ASVs belonging to a phylum but with unassigned genus are grouped together and shown as the lightest shade of the phylum's colour (e.g. unassigned genuses belonging to Firmicutes are shown as the lightest shade of red). Phyla are ordered bottom-to-top

by descending total relative abundance. The 1st colour bar below denotes the location of the subject on the day of sampling, coloured by region. The 2nd colour bar denotes whether the subject self-reported illness on that sampling day.

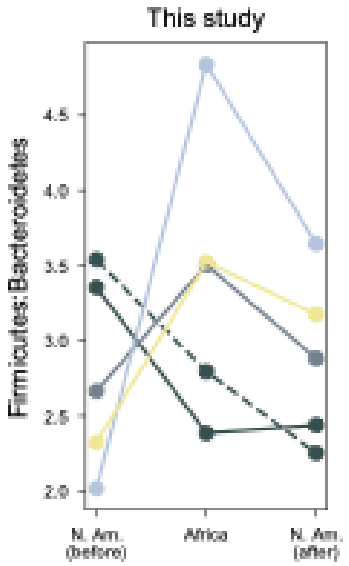


Figure 5. Ratio of mean relative abundances of Firmicutes to Bacteroidetes in our subjects before, during, and after travel to Africa. Dashed line denotes Subject 1's 2nd trip to Africa.

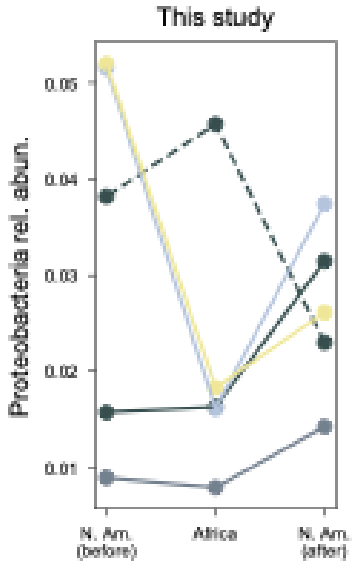


Figure 6. Relative abundance of Proteobacteria in our subjects before, during, and after travel to Africa. Dashed line denotes Subject 1's 2nd trip to Africa.

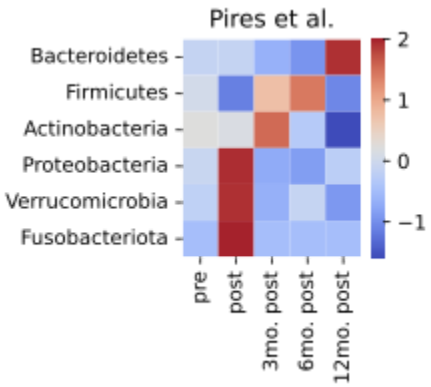
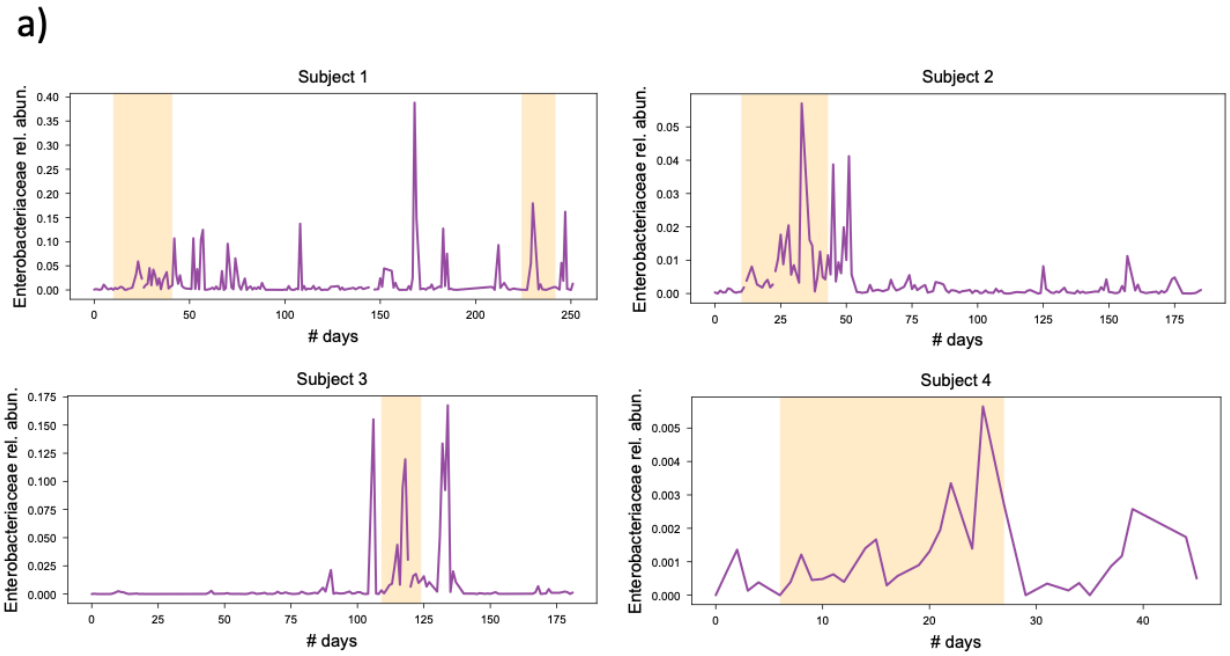


Figure 7. Heatmap of mean relative abundances of most common gut phyla, z-scored across time points in the Pires et al. data.



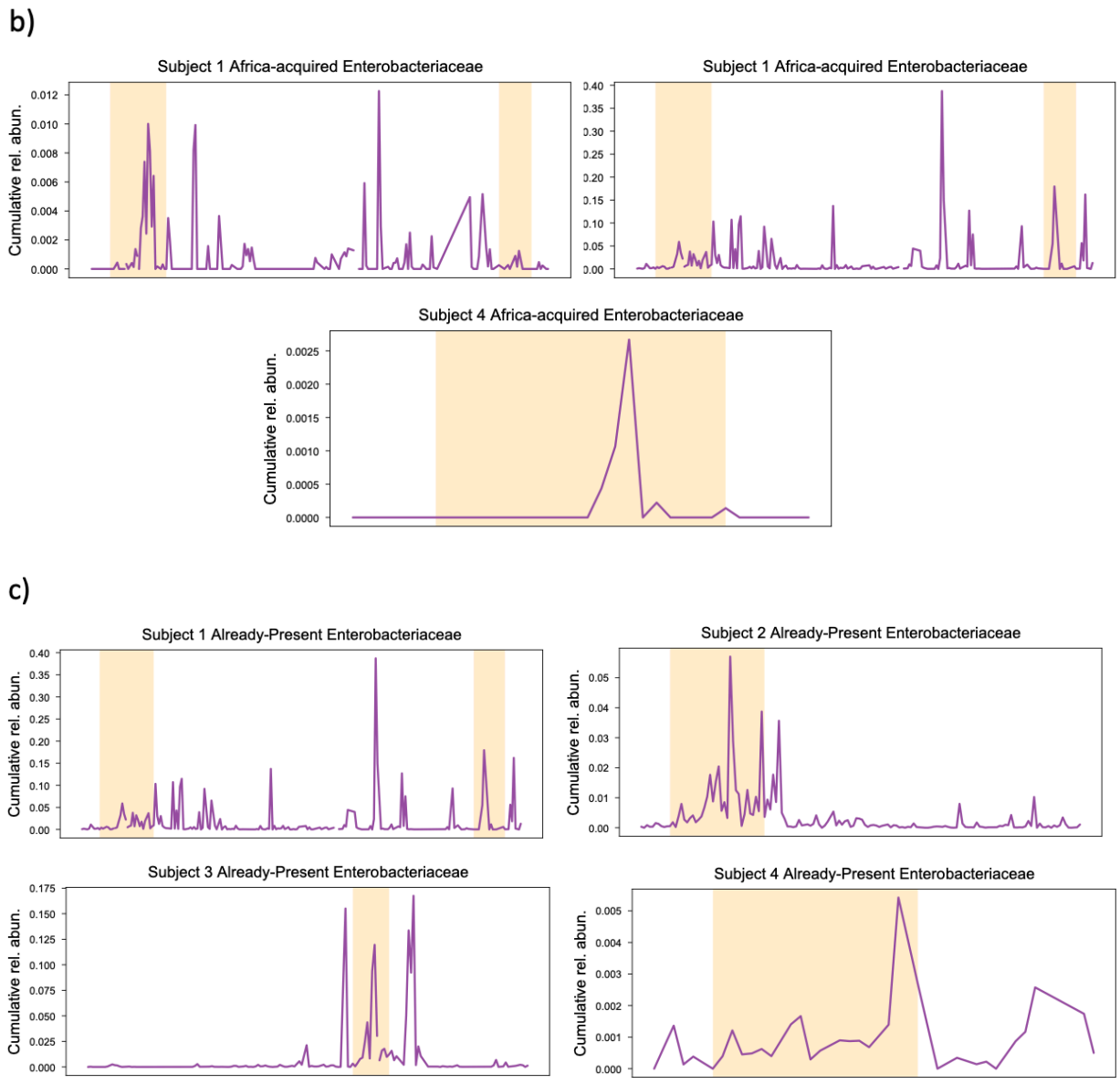


Figure 8. a) Relative abundance of Enterobacteriaceae in our subjects' time series. Periods in Africa are coloured in orange background. b) Relative abundance of Africa-acquired Enterobacteriaceae in our subjects' time series. Subject 3 did not acquire any Enterobacteriaceae while in Africa. c) Relative abundance of Enterobacteriaceae already present pre-African travel in our subjects' time series.

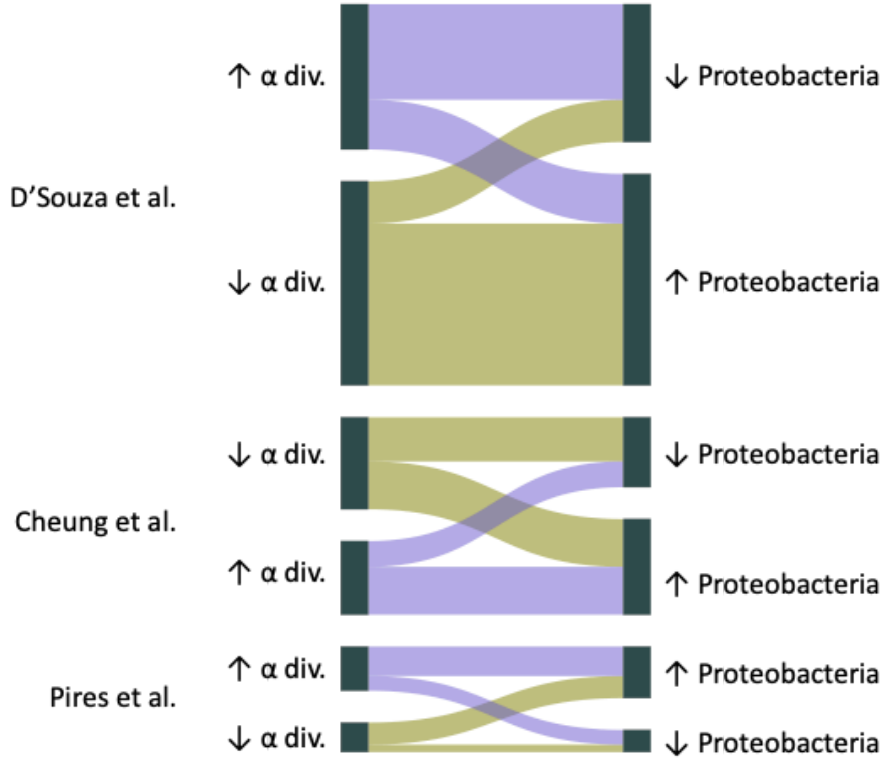


Figure 9. Sankey plots of the association between post-travel outcomes (Δ α diversity, Δ Proteobacteria relative abundance) in the D'Souza et al., Cheung et al., and Pires et al. travelers.

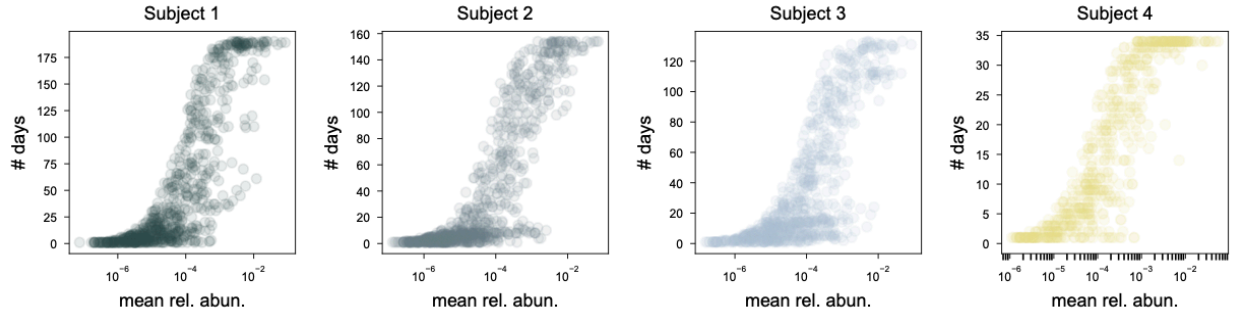
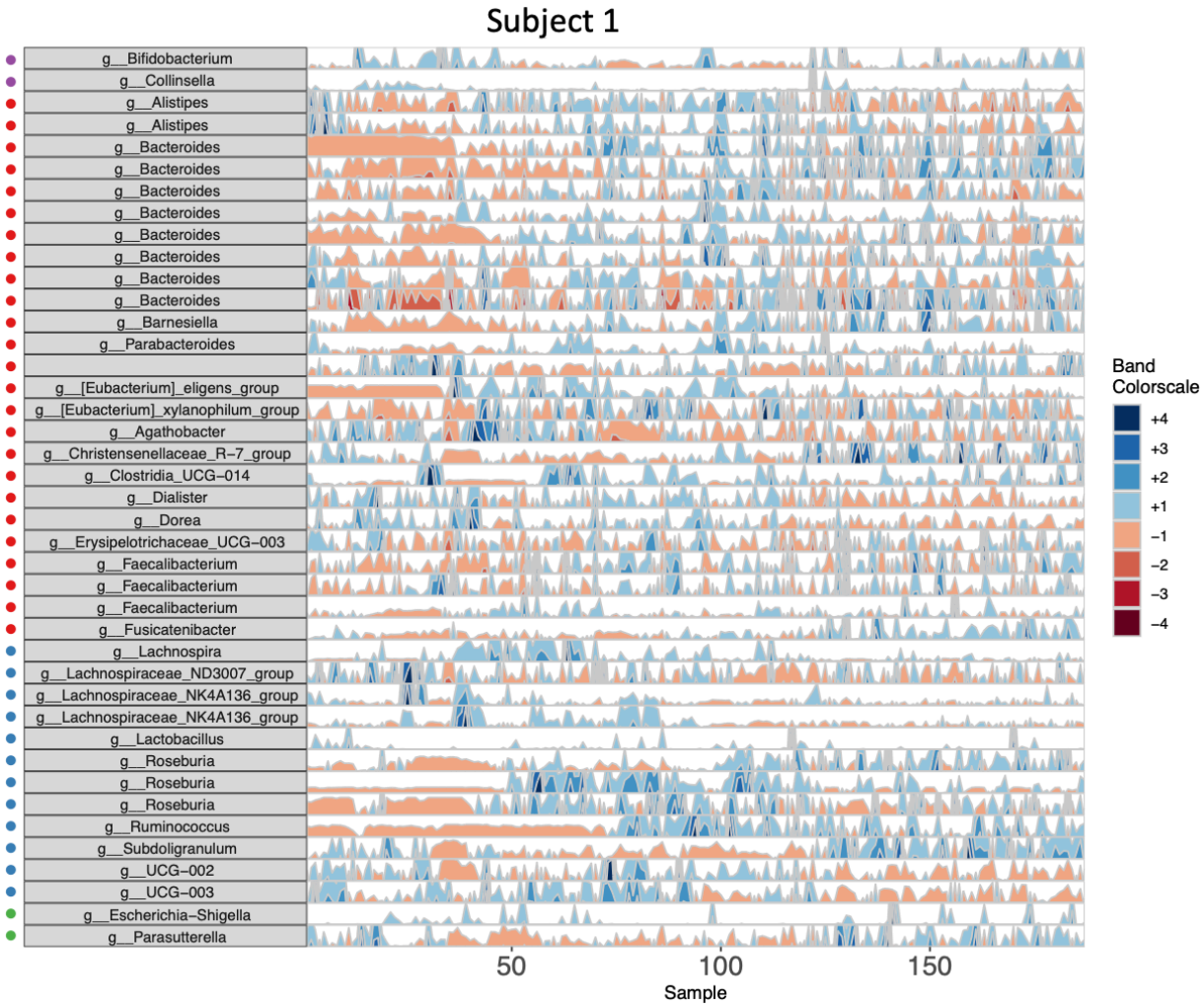
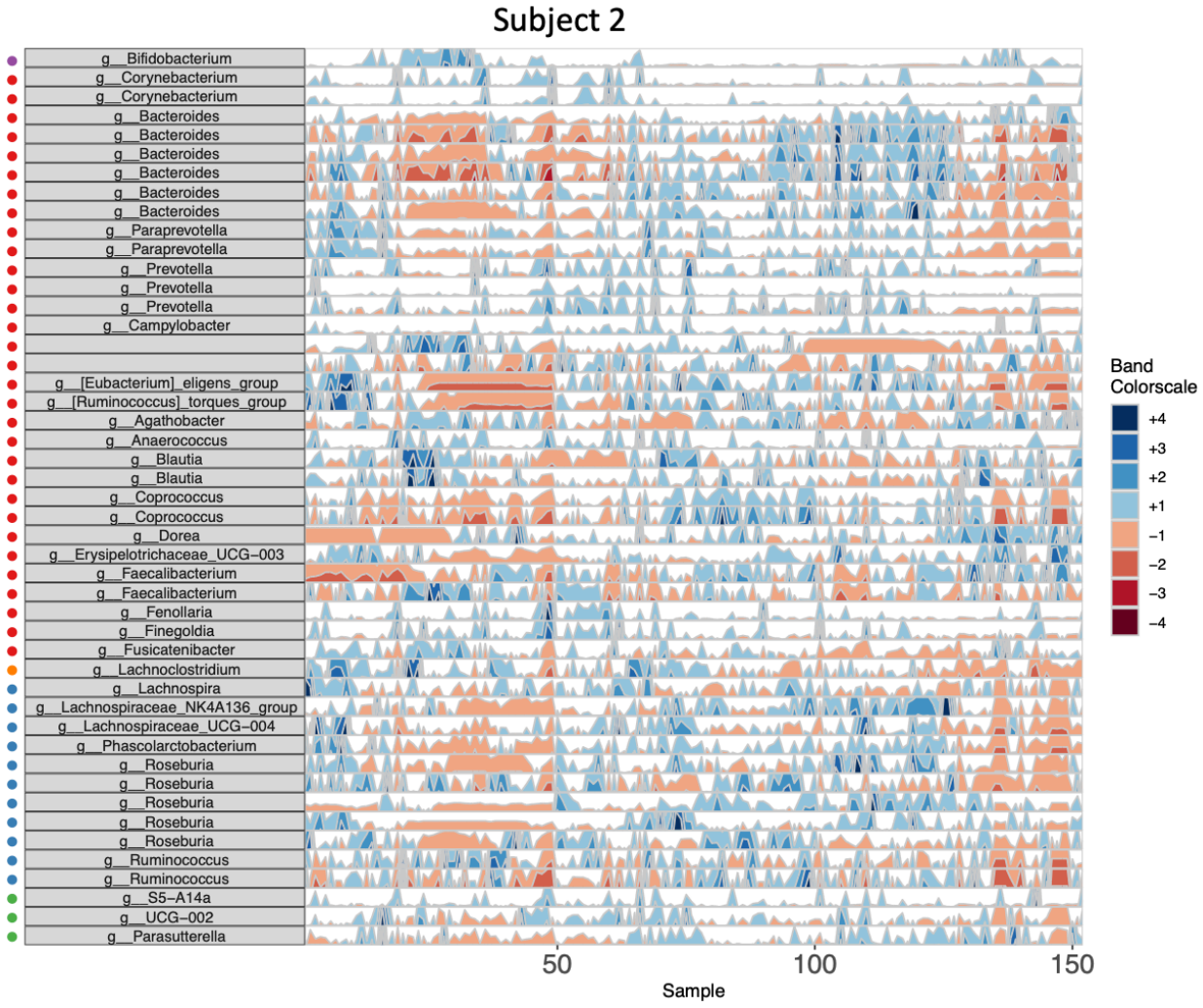
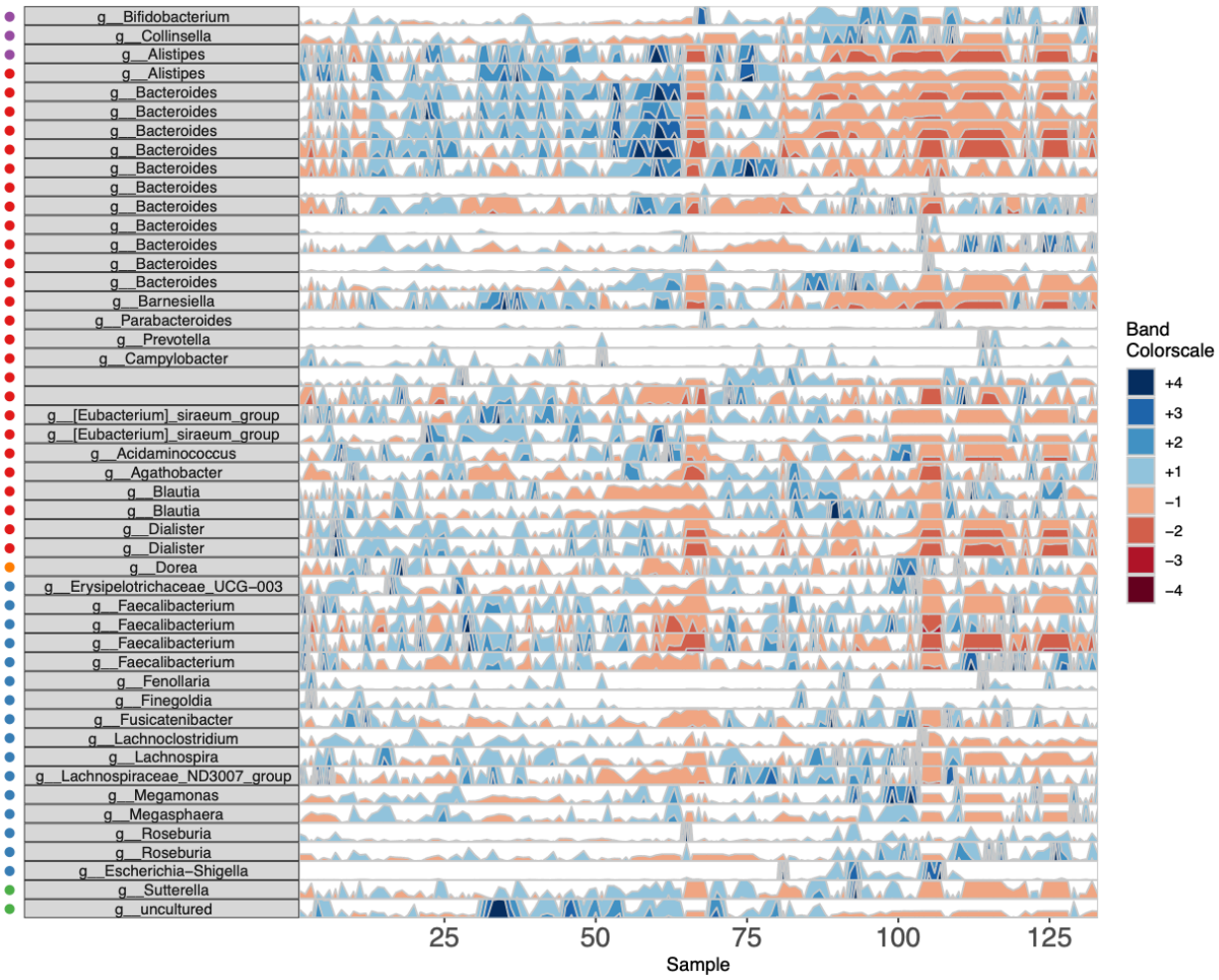


Figure 10. Temporal persistence vs. mean relative abundance of ASVs in each subject's time series.


Subject 1
g__Bifidobacterium
g__Collinsella
g__Alistipes
g__Alistipes
g__Bacteroides
g__Bacteroides
g__Bacteroides
g__Bacteroides
g__Bacteroides
g__Bacteroides
g__Bacteroides
g__Bacteroides
g__Barnesiella
g__Parabacteroides
g__[Eubacterium]_eligens_group
g__[Eubacterium]_xylanophilum_group
g__Agathobacter
g__Christensenellaceae_R-7_group
g__Clostridia_UCG-014
g__Dialister
g__Dorea
g__Erysipelotrichaceae_UCG-003
g__Faecalibacterium
g__Faecalibacterium
g__Faecalibacterium
g__Fusicatenibacter
g__Lachnospira
g__Lachnospiraceae_ND3007_group
g__Lachnospiraceae_NK4A136_group
g__Lachnospiraceae_NK4A136_group
g__Lactobacillus
g__Roseburia
g__Roseburia
g__Roseburia
g__Ruminococcus
g__Subdoligranulum
g__UCG-002
g__UCG-003
g__Escherichia-Shigella
g__Parasutterella
50
100
150
Sample
Band Colorscale
+4
+3
+2
+1
−1
−2
−3
−4


Subject 2
g__Bifidobacterium
g__Corynebacterium
g__Corynebacterium
g__Bacteroides
g__Bacteroides
g__Bacteroides
g__Bacteroides
g__Bacteroides
g__Bacteroides
g__Paraprevotella
g__Paraprevotella
g__Prevotella
g__Prevotella
g__Prevotella
g__Campylobacter
g__[Eubacterium]_eligens_group
g__[Ruminococcus]_torques_group
g__Agathobacter
g__Anaerococcus
g__Blautia
g__Blautia
g__Coprococcus
g__Coprococcus
g__Dorea
g__Erysipelotrichaceae_UCG−003
g__Faecalibacterium
g__Faecalibacterium
g__Fenollaria
g__Finegoldia
g__Fusicatenibacter
g__Lachnoclostridium
g__Lachnospira
g__Lachnospiraceae_NK4A136_group
g__Lachnospiraceae_UCG−004
g__Phascolarctobacterium
g__Roseburia
g__Roseburia
g__Roseburia
g__Roseburia
g__Roseburia
g__Ruminococcus
g__Ruminococcus
g__S5−A14a
g__UCG−002
g__Parasutterella
50
100
150
Sample
Band Colorscale
+4
+3
+2
+1
−1
−2
−3
−4


Subject 3
g__Bifidobacterium
g__Collinsella
g__Alistipes
g__Alistipes
g__Bacteroides
g__Bacteroides
g__Bacteroides
g__Bacteroides
g__Bacteroides
g__Bacteroides
g__Bacteroides
g__Bacteroides
g__Bacteroides
g__Bacteroides
g__Bacteroides
g__Barnesiella
g__Parabacteroides
g__Prevotella
g__Campylobacter
g__[Eubacterium]_siraeum_group
g__[Eubacterium]_siraeum_group
g__Acidaminococcus
g__Agathobacter
g__Blautia
g__Blautia
g__Dialister
g__Dialister
g__Dorea
g__Erysipelotrichaceae_UCG−003
g__Faecalibacterium
g__Faecalibacterium
g__Faecalibacterium
g__Faecalibacterium
g__Fenollaria
g__Finegoldia
g__Fusicatenibacter
g__Lachnoclostridium
g__Lachnospira
g__Lachnospiraceae_ND3007_group
g__Megamonas
g__Megasphaera
g__Roseburia
g__Roseburia
g__Escherichia−Shigella
g__Sutterella
g__uncultured
25
50
75
100
125
Sample
Band Colorscale
+4
+3
+2
+1
−1
−2
−3
−4

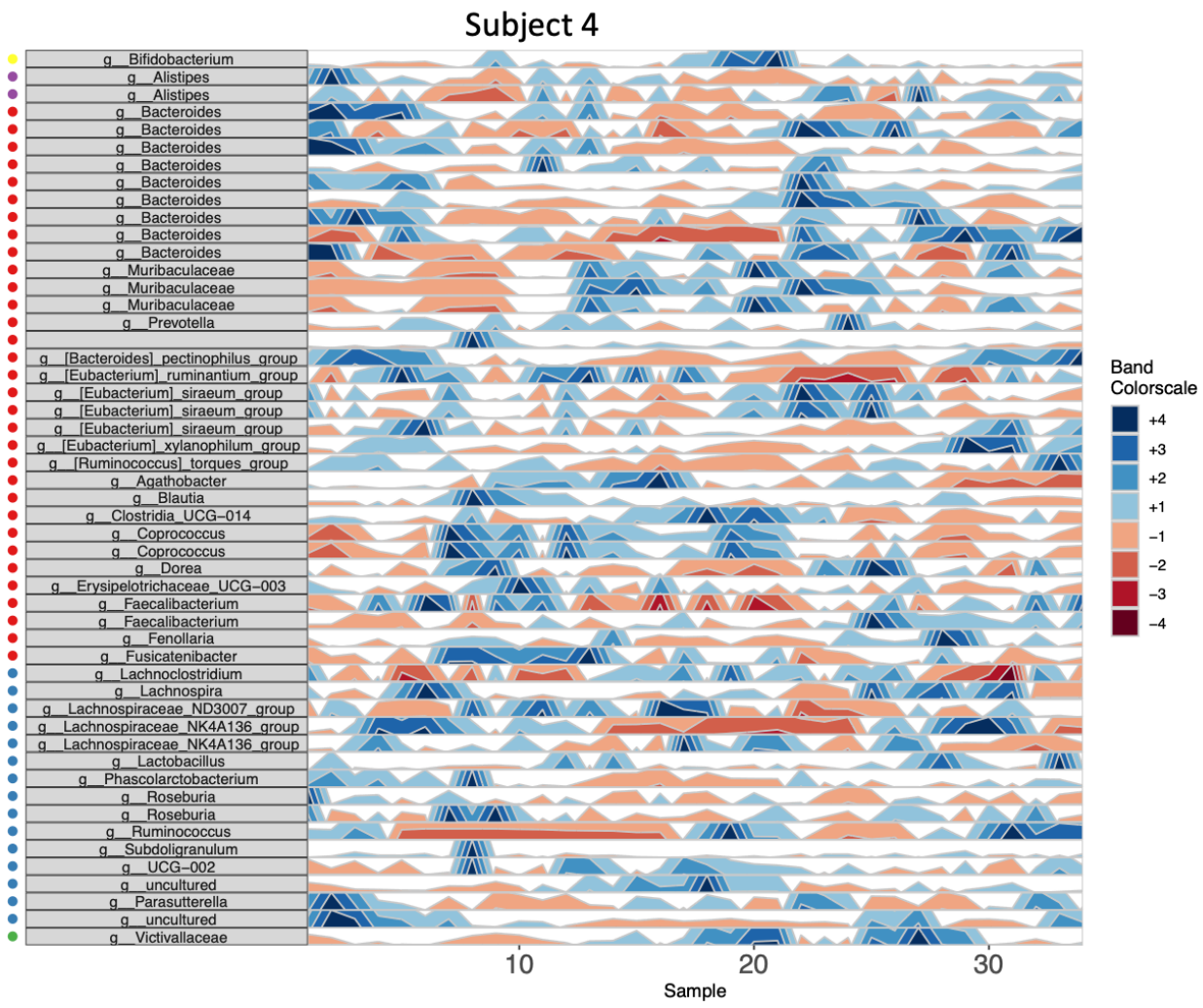


Figure 11. Horizon plots depicting abundance fluctuations of the most abundant and prevalent ASVs in each subject's time series. Abundance values are centred at zero (indicating the median relative abundance for that ASV) and colours reflect the quartile abundance above (blue) or below (red) the median. Dots on the left denote the phylum the ASV belongs to (with the same colour scheme as Figure 4). ASVs are labeled with their genus classifications.

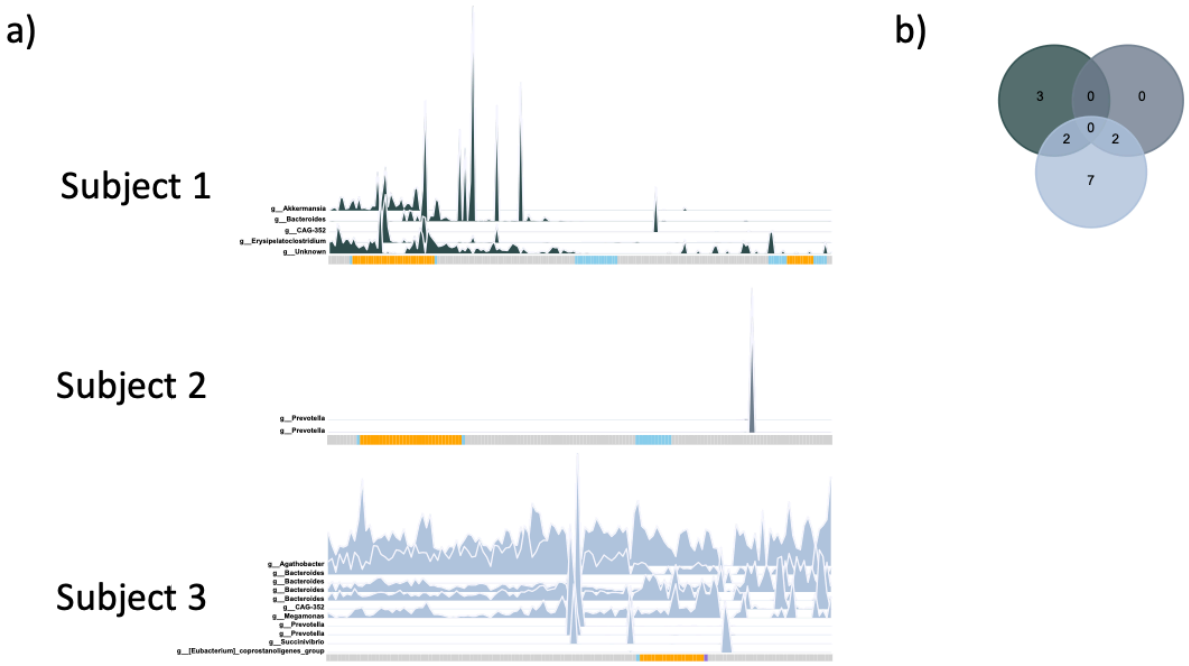


Figure 12. a) Conditionally-rare taxa abundances through subject time series. CRTs were identified as ASVs with Sarle's bimodality coefficient > 0.8 and peak relative abundance ≥ 10%. b) Venn diagram of shared CRTs among subjects. No CRTs were identified in Subject 4's time series.

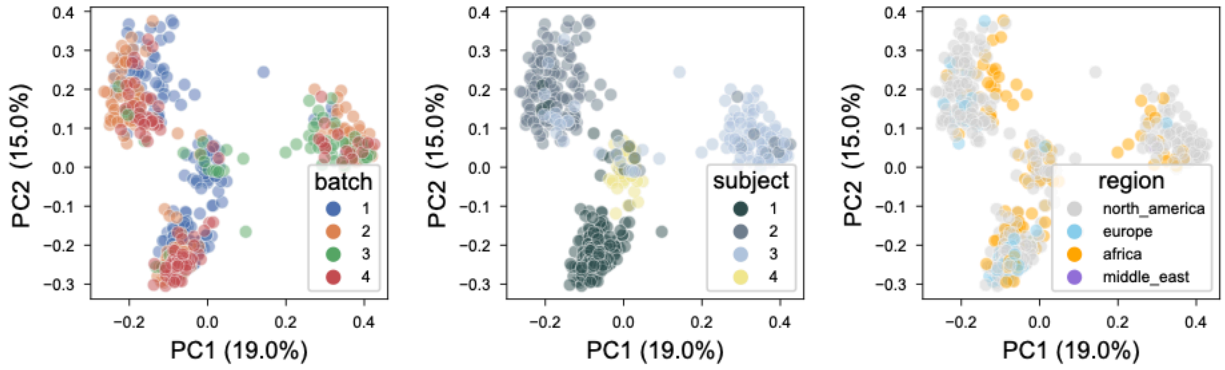


Figure 13. PCoA plots of samples in all of our subjects' time series, coloured by sequencing batch, subject, and geographic region.

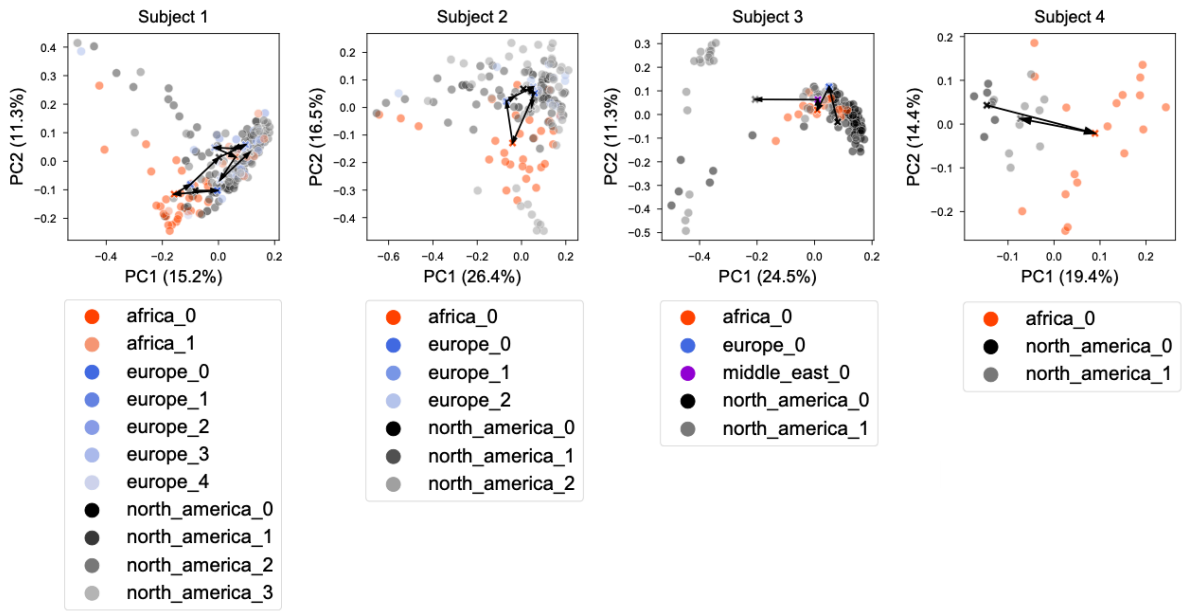


Figure 14. PCoA plots of each subject's samples, coloured by geographic region. Continuous stays in each region are numbered in order (e.g. north_america_0 is a subject's first continuous period in North America, north_america_1 is the period when they next return to North America, etc.) and coloured by geographical region (Africa: shades of red-orange, Europe: shades of blue, North America: shades of black-grey, Middle East: purple). X's denote centroids (e.g. a black X denotes the centroid of all samples from north_america_0 for that subject). Arrows connect consecutive centroids in each time series.

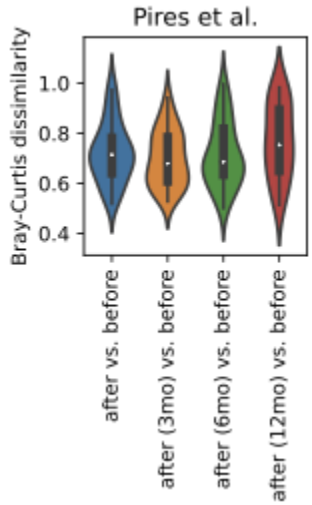


Figure 15. Microbiome divergence (as measured by Bray-Curtis dissimilarity) over time from pre-travel baseline for each subject in the Pires et al. cohort.

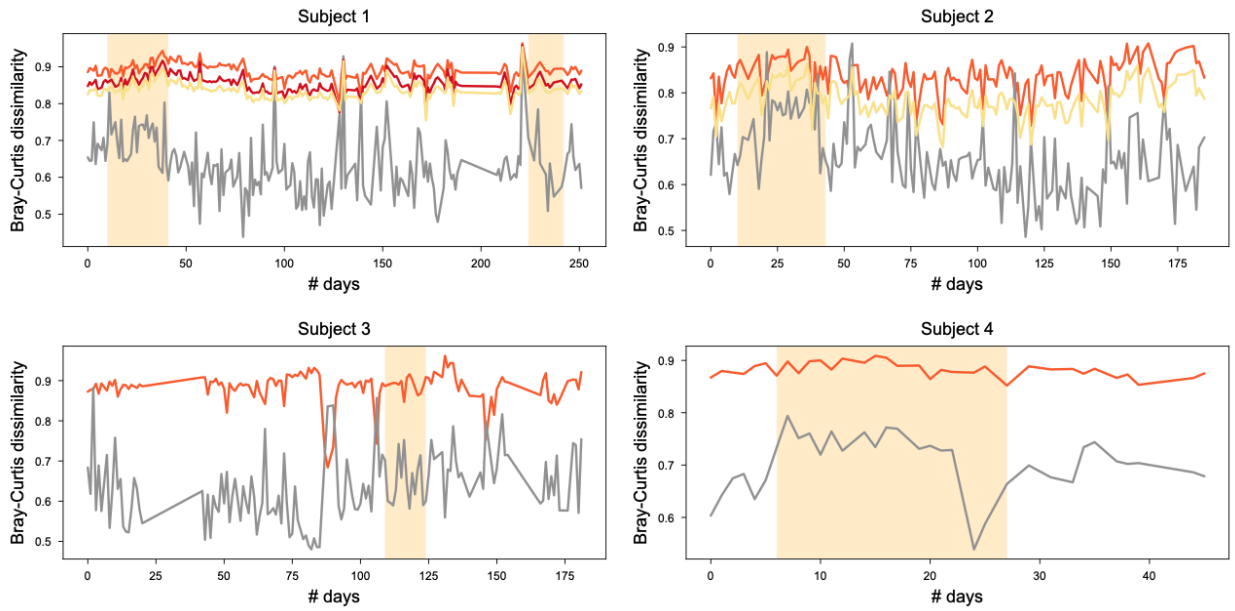


Figure 16. Beta diversity trajectories (as measured by mean Bray-Curtis dissimilarity) from location-matched cross-sectional microbiome sampling cohorts in the US (grey), Rwanda (dark red), Tanzania (orange), and Cameroon (yellow) for each subject's microbiome. Periods in Africa are coloured in orange background.

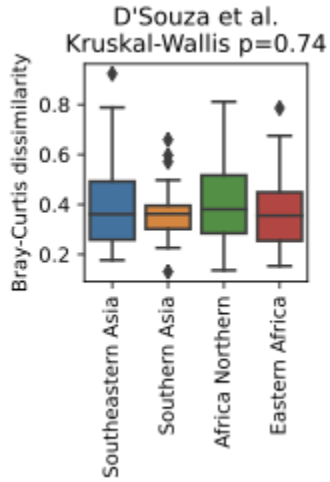


Figure 17. Within-subject divergence from baseline pre-travel (as measured by Bray-Curtis dissimilarity) grouped by destination, among the D'Souza et al. cohort.

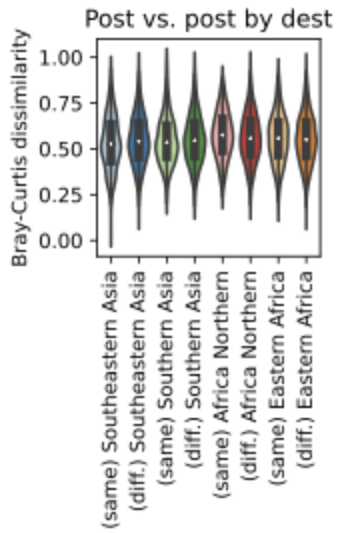


Figure 18. Cross-subject all-versus-all microbiome dissimilarity of post-travel samples (as measured by Bray-Curtis dissimilarity), grouped by shared vs. different destinations, in the D'Souza et al. cohort.

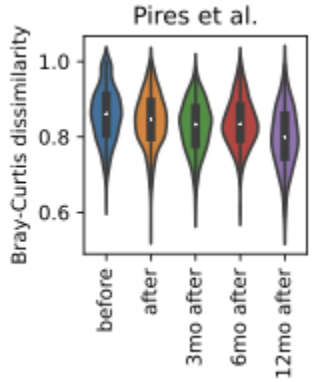


Figure 19. Cross-subject all-versus-all microbiome dissimilarity, grouped by time point, in the Pires et al. cohort.

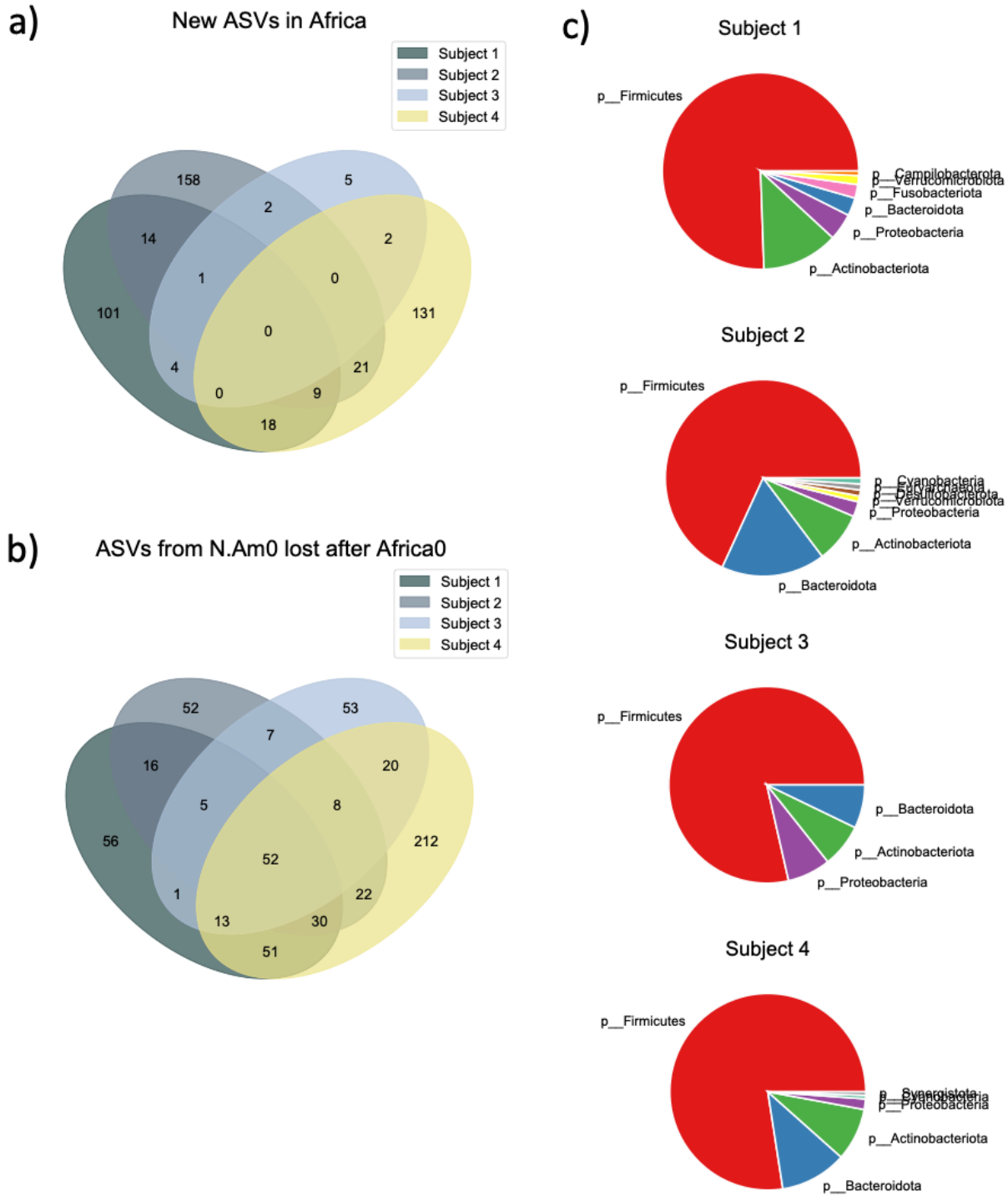


Figure 20. a) Venn diagram of newly-colonizing ASVs in Africa in common between subjects. For Subject 1, only the 1st trip to Africa is included. b) Venn diagram of ASVs that disappeared in Africa in common between subjects. For Subject 1, only the 1st trip

to Africa is included. c) Pie charts of phylum assignments of all newly-colonizing ASVs in Africa for each subject.

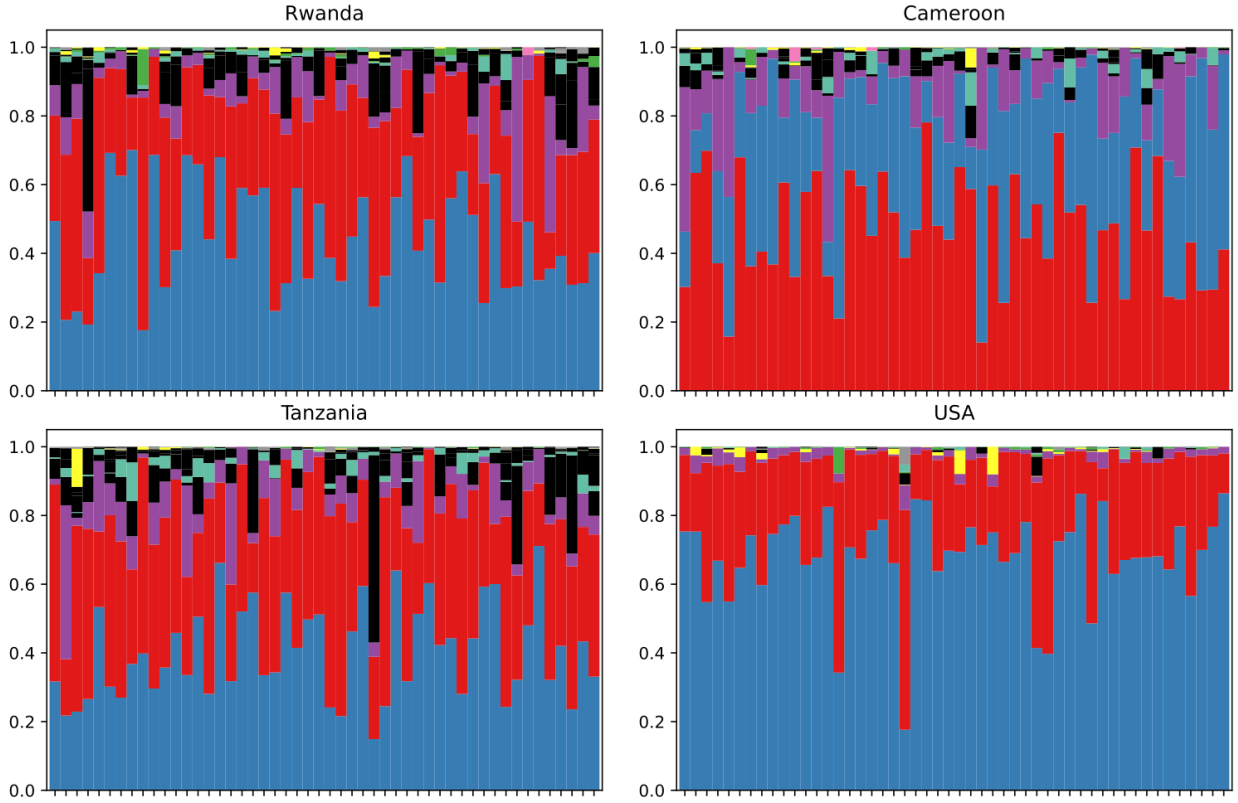


Figure 21. Microbiome composition of 50 example subjects from each of Rwanda, Cameroon, Tanzania, and USA in the cross-sectional dataset. Microbes are coloured by phylum according to the same colour scheme as Figure 4, with black being assigned to all phyla which were never detected in our travelers' gut samples.

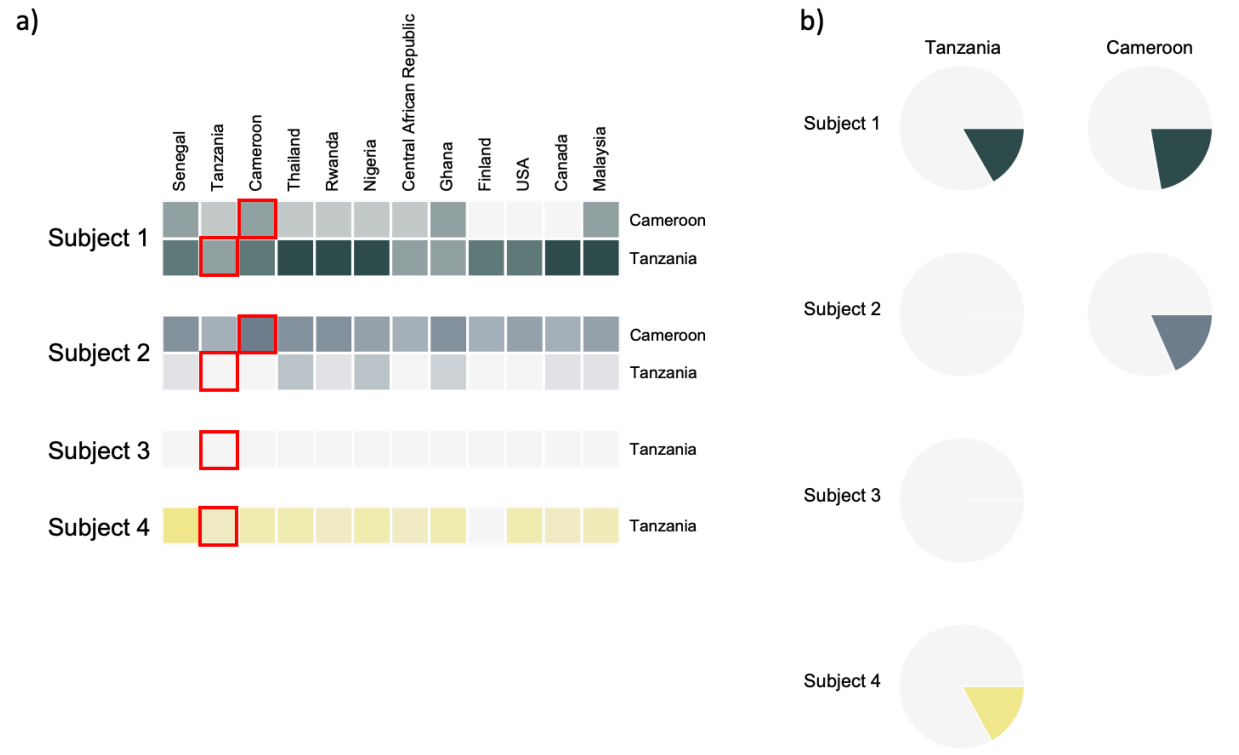


Figure 22. a) Overlap of genuses newly-appearing in each African country with those found in country-matched local microbiome sampling cohorts (highlighted in red). A darker colour denotes greater overlap (more genuses in common between travelers and locals in that country). b) Fraction of newly-colonizing genuses in Africa that were shared with country-matched local cohort.

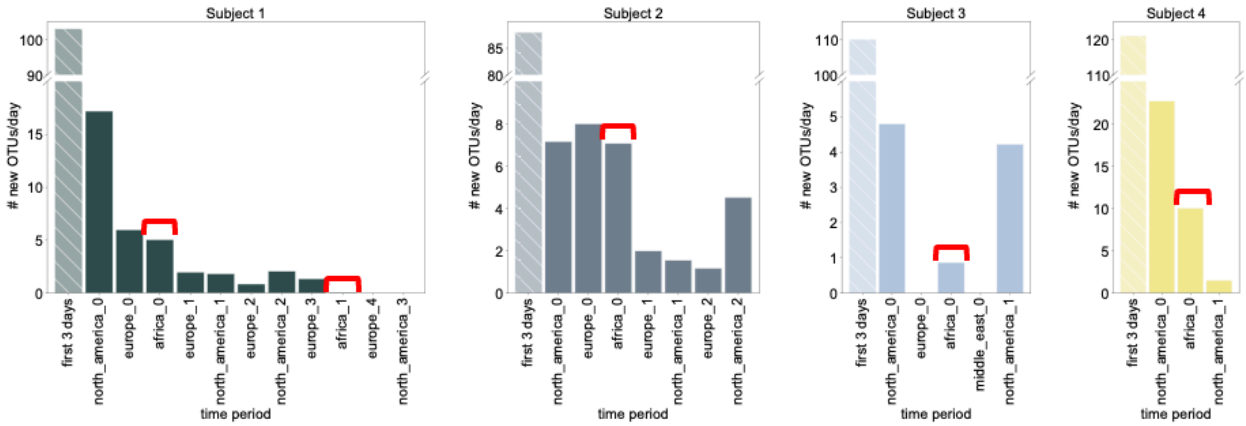


Figure 23. Mean number of newly-appearing ASVs per day in each continuous time period in each geography. The first 3 days of each subject time series is represented as

a separate time period, to account for core or already-present ASVs (those that are not actually new). In-Africa periods highlighted in red.

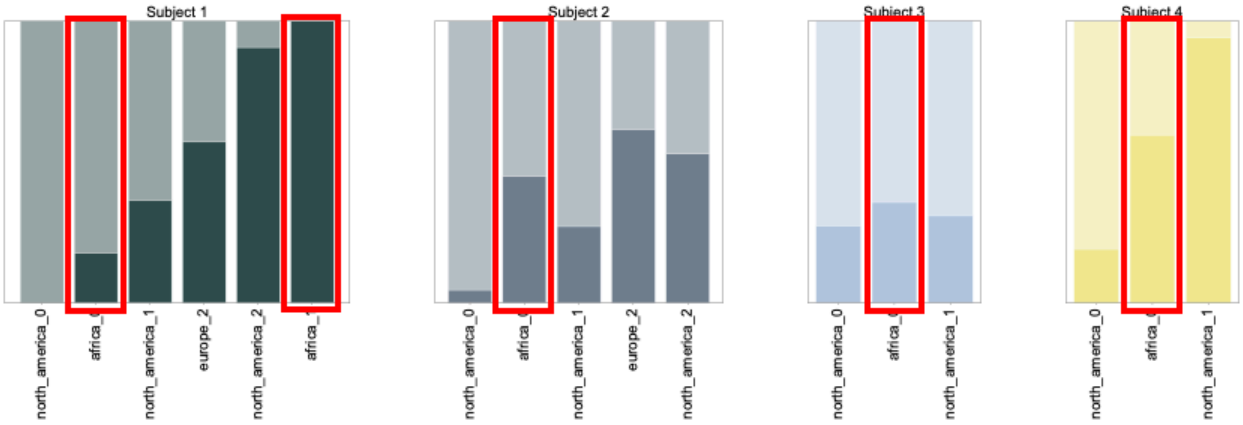


Figure 24. Fraction of newly-appearing ASVs in each longer continuous time period in each geography which exhibited high (>3 sampling days, pale colour) or low (≤3 sampling days, opaque colour) temporal persistence. In-Africa periods highlighted in red.

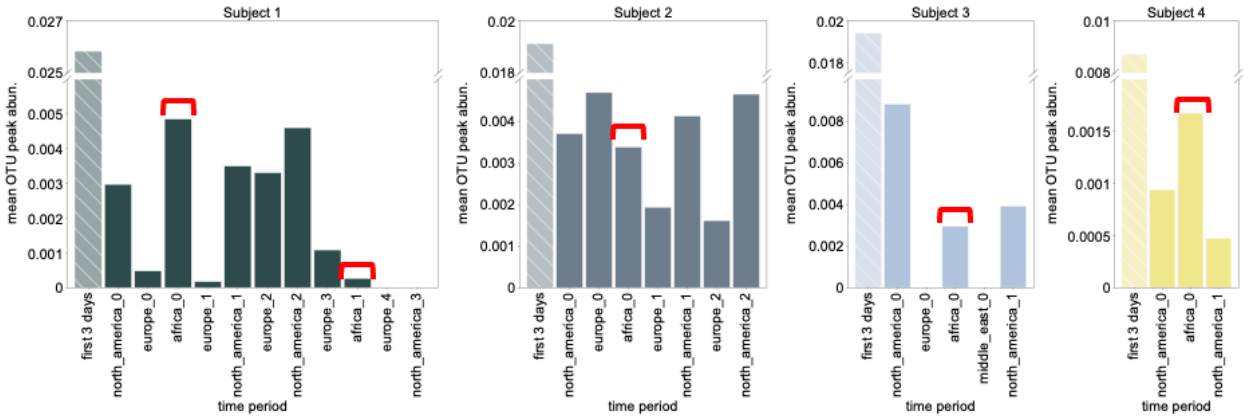


Figure 25. Mean peak relative abundance of new ASVs in each continuous time period in each geography. The first 3 days of each subject time series is represented as a separate time period, to account for core or already-present ASVs (those that are not actually new). In-Africa periods highlighted in red.

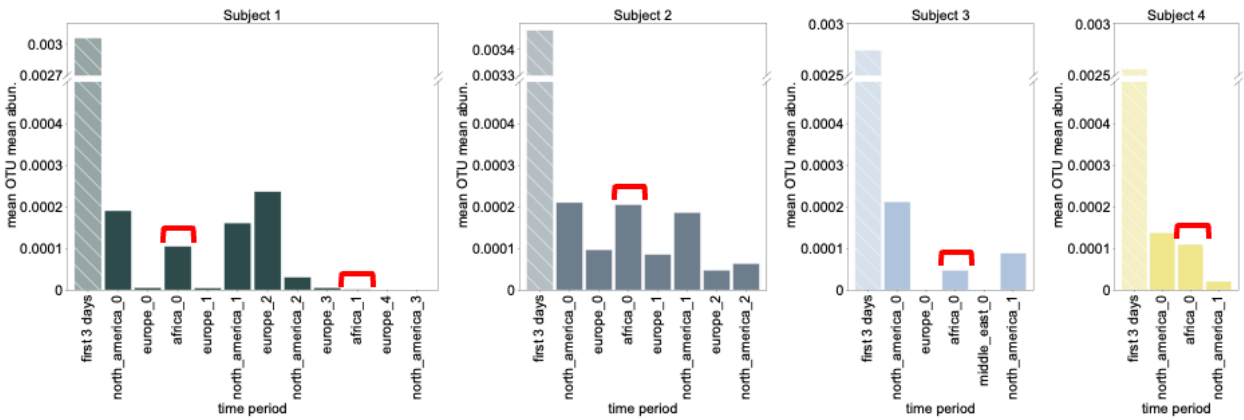


Figure 26. Mean (across ASVs) of mean relative abundance (across time) of each new ASVs in each continuous time period in each geography. The first 3 days of each subject time series is represented as a separate time period, to account for core or already-present ASVs (those that are not actually new). In-Africa periods highlighted in red.

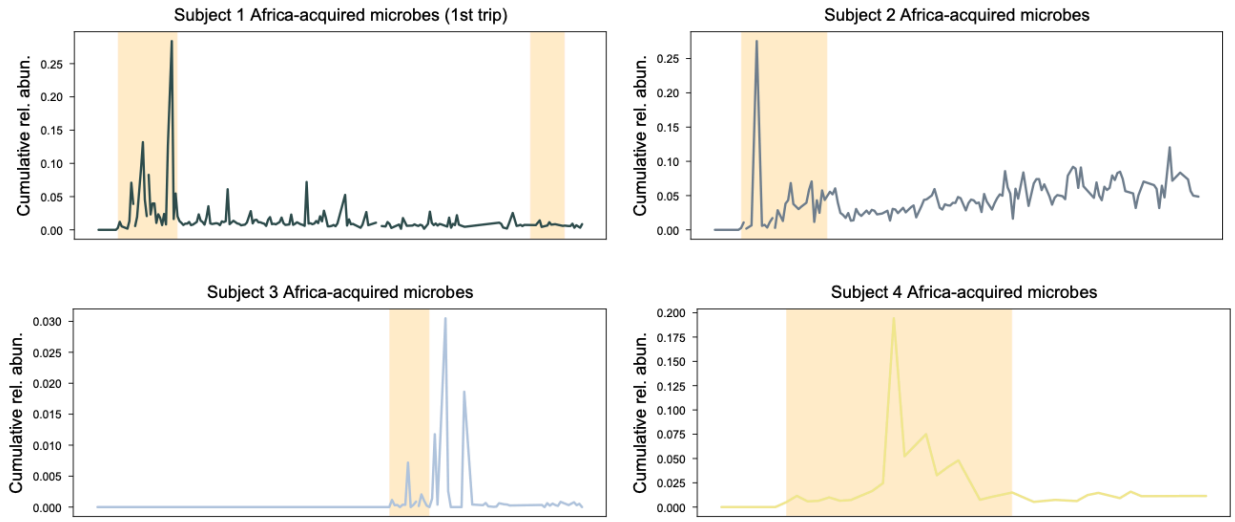


Figure 27. Total relative abundance of Africa-acquired microbes over time. Only microbes acquired during Subject 1's first trip to Africa are shown. Periods in Africa are coloured in orange background.

Chapter 2: Strain-level diversity of antiphage defense capabilities across gut bacterial phyla

This chapter was a collaboration between me and Jay Zhao. Jay and his undergraduate mentees generated StrainAtlas biobank isolates and genome data, I performed the antiphage analyses, and Jay and I performed visualization. This chapter is written by me.

Background

The typical human gut microbiome comprises hundreds to thousands of distinct bacterial species (Costea et al.). Beyond this inter-species-level diversity, it is known that different human hosts can harbour different strains of a given species, and that one host microbiome can contain multiple strains of the same species, coexisting. Research has uncovered vast strain-to-strain phenotypic variability among many gut bacteria, with many of these findings carrying significant relevance to host health. For example, it is widely known that antibiotic resistance capacity within a species can range from susceptible to fully resistant. This can be seen in the classical case of *Escherichia coli*, which naively is susceptible to most clinical antibiotics, but in practise is able to acquire any of a wide arsenal of resistance via its neighbours in addition to its own mutational processes (Poirel et al.). Pathogen virulence can vary similarly highly within species. As well, it has been observed more recently that there can be differing host immune responses to different strains of the same gut species, with one group observing a range of IgA secretion levels by identical hosts in response to different *Bacteroides ovatus* strains (Yang et al.). Researchers have also shown that there can be strain-level differences relevant to host therapeutics, such as bacterial metabolism of host-administered drugs and anti-cancer activity (Chambers and Illingworth).

Collectively, this observed functional diversity at the strain level is underpinned by the vast genomic diversity of gut bacterial species. To grapple with the scale of this diversity, researchers have introduced the concept of the pangenome to describe the whole set of genes in a species (Tettelin et al.). Alongside this are the related concepts of the core genome (the subset of genes in the pangenome shared by most or all species members, often essential or housekeeping genes) and the accessory genome (the subset of genes shared by only some species members and more likely to be dispensable). Some researchers break down the accessory genome further into shell and cloud sections (medium- and low-prevalence genes, respectively) (Wolf et al.). In many gut species, the size of the pangenome greatly outmeasures the size of the core genome or any individual strain genome. This indicates that the pool of possible gene content (and thus functional capacity) of the species as a whole is far greater than can

be captured in any single genome. As an example, *E. coli* is estimated to have a core genome of 2000-3000 genes, but by comparison its pangenome is estimated at 100,000+ genes (Park et al.). Furthermore, the average *E. coli* genome only contains about 5000 genes total. Thus, to fully grasp the scope of genetic and functional diversity of *E. coli*, one needs to study far more than a single *E. coli* genome.

The same is true for other species in the gut, but while *E. coli* is among the best-studied and most-sequenced species, for many others there is a paucity of publicly-available, high-quality sequenced genomes or cultured strains (Blackwell et al.).

To address this gap in data availability, our lab has developed a rich biobank and data resource called StrainAtlas, which comprises 15,000+ cultured gut bacterial strains with sequenced whole genomes from 700+ human donors from varied geographies and demographics. This new resource differs from previously-released biobank resources in its uniquely-deep intra-species sampling, and thus it is ideal for exploration of strain-level diversity.

Among the many little-studied questions in this domain of strain-level diversity include that of strain-to-strain differences in microbial anti-bacteriophage defense capabilities. A recently-burgeoning area of study for microbiologists broadly, antiphage mechanisms are now known to be far more diverse than previously believed, in good part because of a landmark study in 2018 which discovered ten new defense systems through an exploratory computational approach with experimental validation, which leveraged their known patterns of genomic colocation (Doron et al.). The authors used 45,000+ bacterial and archaeal genomes for their initial search, which comprised the entirety of public genomes on the NCBI at the time.

Since then and with the renewed interest in this area, researchers have unearthed more novel defense systems (Gao et al.; Bernheim et al.; Millman et al.), and automated their computational annotation in genome assemblies (Tesson et al.; Payne et al.). Building upon this and with the vast new resource of StrainAtlas, we demonstrate the extent of strain-level diversity in antiphage defense capabilities across the phylogenetic tree of gut bacteria. Our results point toward there being much more to discover in the way of

antiphage defense mechanisms in model and non-model bacteria and underscore the value and importance of resources like StrainAtlas in capturing strain heterogeneity.

Methods

In brief, stool samples from human donors were plated and grown on selective media, and colonies were picked for species-level identification via 16S amplicon sequencing. Colonies were then restreaked, grown, and arrayed onto 96-well plates for Illumina whole genome sequencing. Whole genomes were then assembled from FASTQ reads using Unicycler (Wick et al.), and assigned species identifiers using rMLST (Jolley et al.).

Genomes were filtered for inclusion in the phage defense analysis by average read depth (minimum=5), and genome FASTA files in nucleotide format were translated to amino acid format with Prodigal (Hyatt et al.).

Amino acid files were then inputted for phage defense system annotation into DefenseFinder (Tesson et al.). Briefly, DefenseFinder employs a two-step method for prokaryotic antiviral system identification: first, it uses a Hidden Markov Model-based sequence homology detection step to find individual sequences resembling those of known defense systems in the literature (manually curated by the authors); and second, it checks that the presence and arrangement of these sequences abide by the set of “decision rules” for the system containing that sequence, again prescribed by the authors and based on its genetic architecture as previously studied. The HMM database used was up-to-date as of December 2022, and the software was run with default (conservative) settings. Results were analyzed and visualized in Python.

Results

An overview of the antiphage arsenal in gut bacteria

In total, 16,610 StrainAtlas genomes are included in these analyses, comprising 181 species and spanning all four major gut phyla. We detect 156,769 different antiphage

systems total, comprising 124 distinct families (e.g. CBASS), 188 subfamilies (e.g. CBASS type I, II), and 365,467 genes. Antiphage system count per genome varies between zero (occurring in 25 genomes) and 46 (found in one genome of *Parabacteroides distasonis*) (**Figure 1a**). Most genomes (18/25) lacking any defense systems belong to just two species, *Latilactobacillus sakei* and *Enterococcus faecium*. Overall, 50% of genomes harbour at least 9 defense systems.

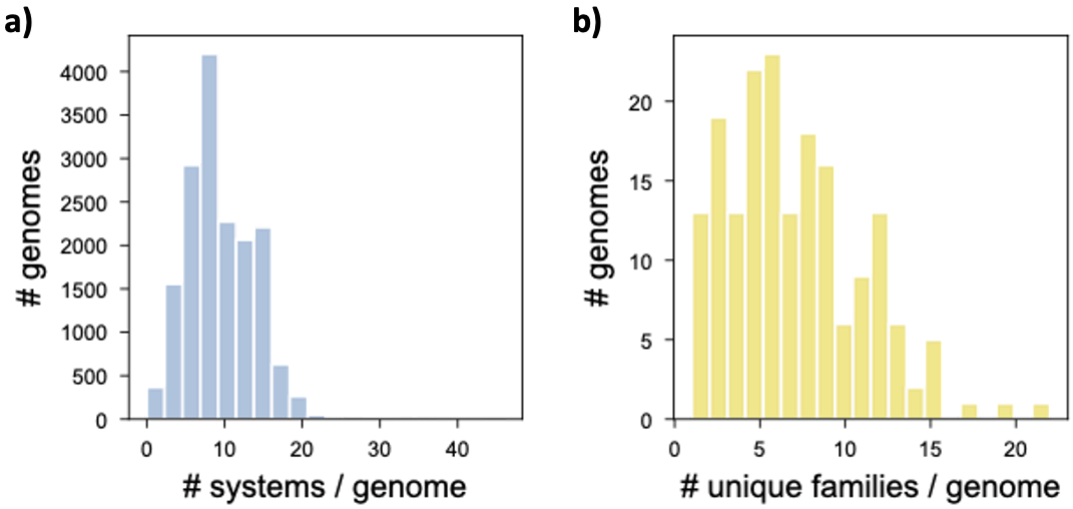


Figure 1. a) Most genomes contain a modest number of total antiphage defense families. b) Similarly for unique defense system families.

To understand the diversity of antiphage defense in StrainAtlas, we calculate the per-genome count of distinct antiphage system subfamilies (which we refer to as just “families” from here on for ease, superseding the previous more general meaning), which is on average 7.24 (**Figure 1b**). Genomes can carry multiple systems of the same family, for example many *Lactobacillus paragasseri* genomes contain multiple Restriction-Modification (RM) type I systems. We observe that generally antiphage system count correlates positively with family count (Pearson r=0.96, p<0.001) (**Figure 2**), suggesting that as the sheer size of antiphage arsenal increases, so too does its diversity. Some genomes are exceptions to this, such as one strain of *Phocaeicola vulgatus*, which has high redundancy among its systems (23 total, 13 unique). Consistent with past findings in the literature, these genomes tend to contain many restriction-modification (RM) systems (Tesson et al.). In fact, we find that for genomes

encoding >5 systems but with high redundancy (# unique systems / # total systems ≤60%, n=87), RM systems comprise 57% of all systems.

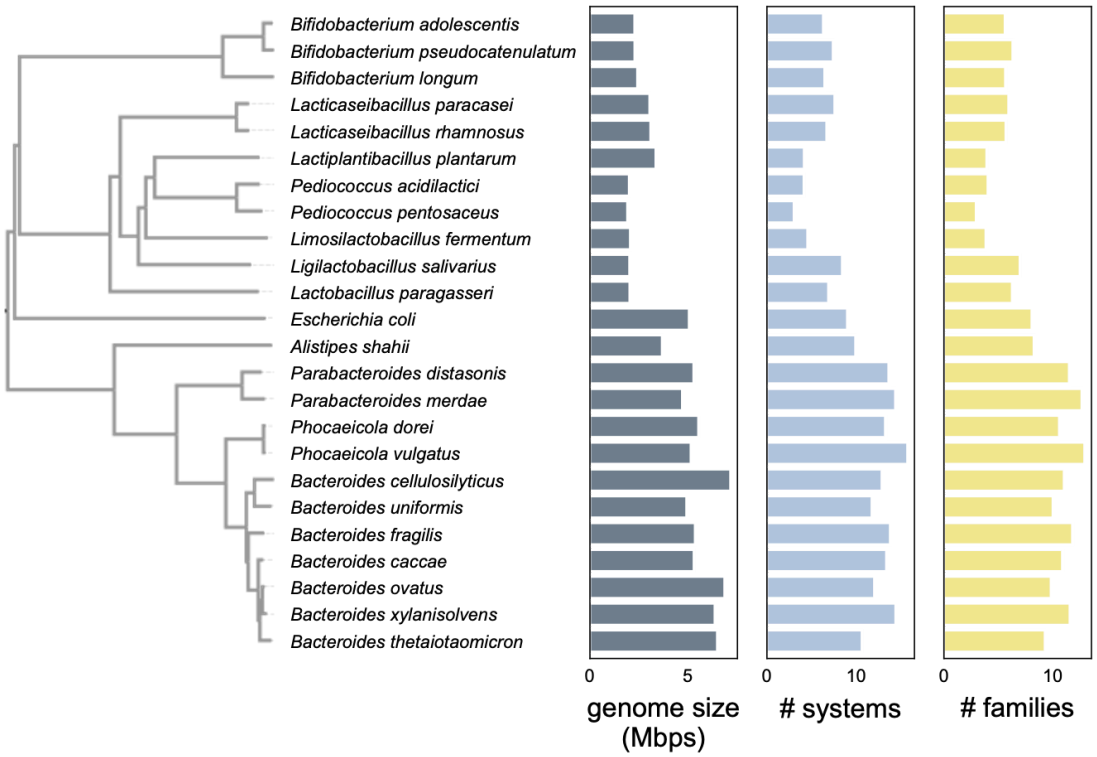


Figure 2. Antiphage systems are found in many gut species (shown are species with ≥100 genomes in StrainAtlas). There are consistent positive correlations between genome size, antiphage system count per genome, and family count per genome between species.

Across and within species, we observe consistent positive correlations between genome size and antiphage system count (**Figure 3a-c**). This makes sense, as genome size reflects accessory system capacity.

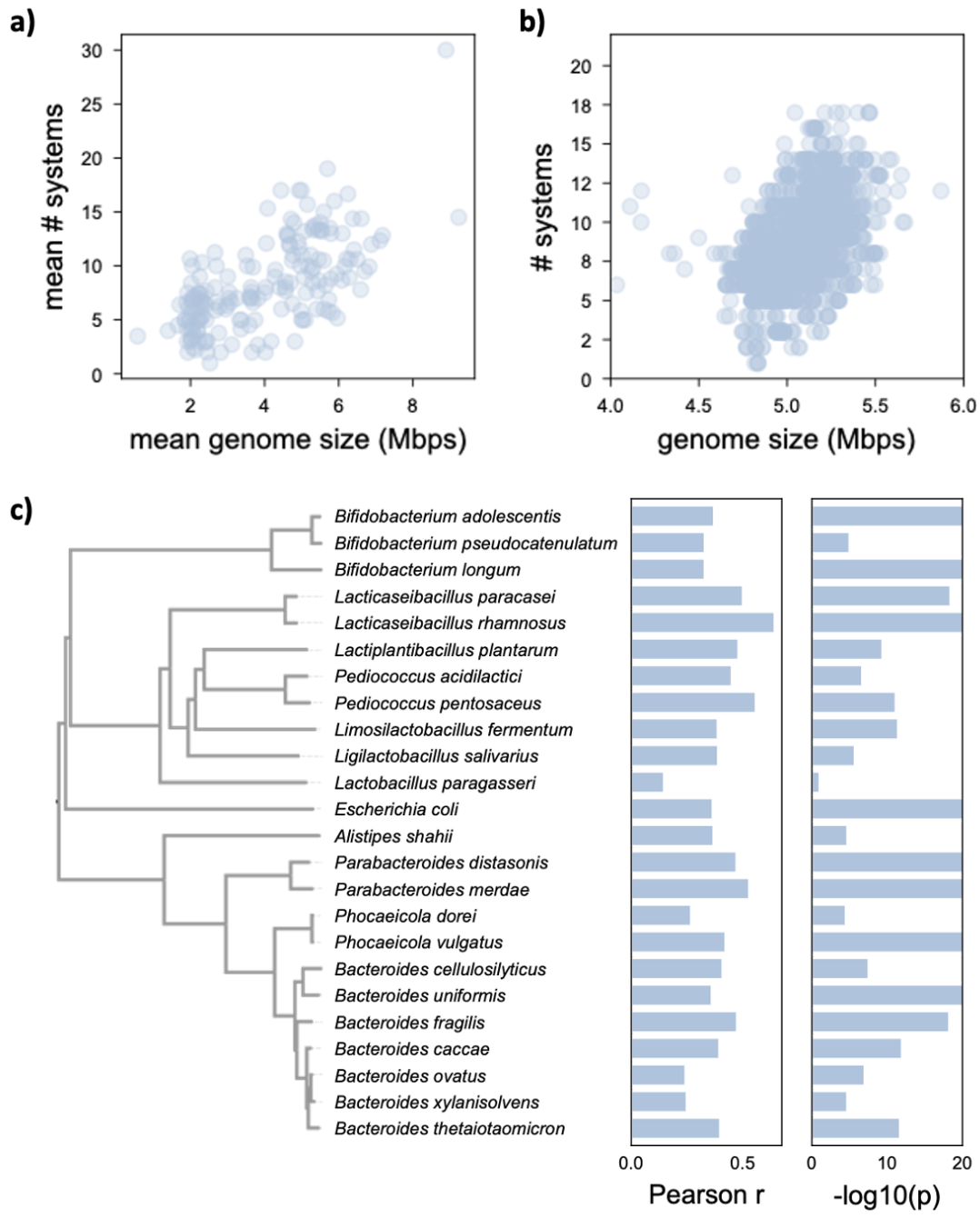


Figure 3. a) Antiphage system abundance positively correlates with genome size across species. b) Antiphage system abundance positively correlates with genome size within a species (*E. coli* is shown). c) This within-species correlation is true for many species (shown are StrainAtlas species with ≥100 genomes).

The antiphage arsenal of gut bacteria is variable across most phyla but enriched in Proteobacteria

We next focus on the antiphage defense families found in StrainAtlas. Among those detected, earlier discovered families are the most predominant, with RM systems showing highest prevalence at 96% of all genomes and CRISPR-Cas systems second at 57% of all genomes (**Figure 4**). These and the handful of other more prevalent families tend also to be the most abundant. Frequencies for remaining detected families range from a single genome (the TerY-phosphorylation triad family) to 46% of genomes (the PARIS abortive infection system family), though most are below 30% prevalence. Among all families detected, 48% are very rare (n=59), found in ≤1% of genomes, and of these 20% are found only in a single species (n=12). Some families are enriched in or restricted to certain phyla. This can be seen in the occurrence patterns of the abortive infection systems AbiG and AbiN, both found across genera within Firmicutes. The most striking example, however, is concentration of families in Proteobacteria. Despite making up only a minority of all strains in StrainAtlas (**Figure 5)**, 80% of all detected families are detected in Proteobacteria genomes (**Figure 6**). This is mostly driven by the outsized enrichment of families in *E. coli*. We posit that this is due not to any unusual intrinsic quality of *E. coli* itself, but rather its history as the best-studied model bacterium. This likely biases the detection method we used (and realistically, any computational detection method) toward *E. coli*-based systems as they would be highly-represented in the methods' databases. Indeed, many of these systems were first discovered in *E. coli* experiments.

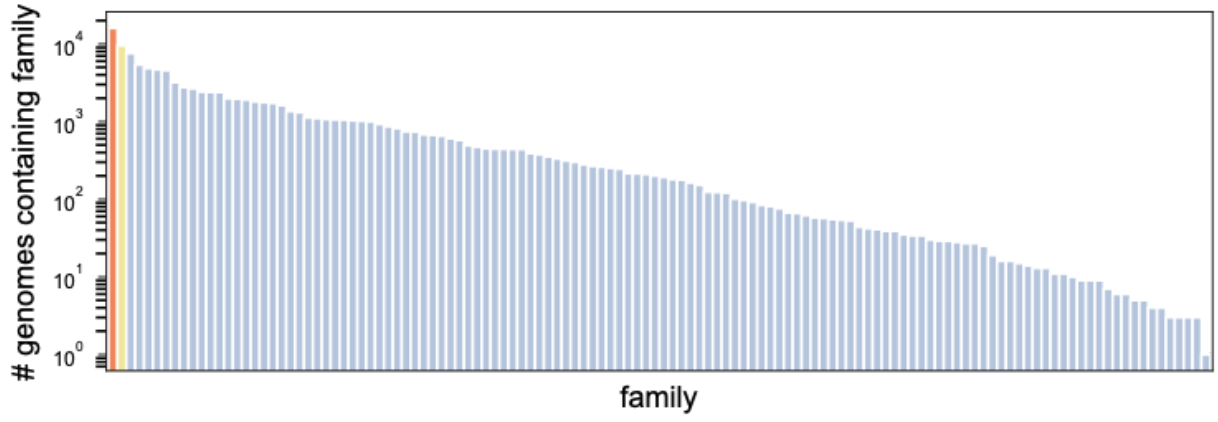


Figure 4. StrainAtlas contains both a vast number and variety of antiphage defense systems, dominated by RM (red) and CRISPR-Cas (yellow) systems.

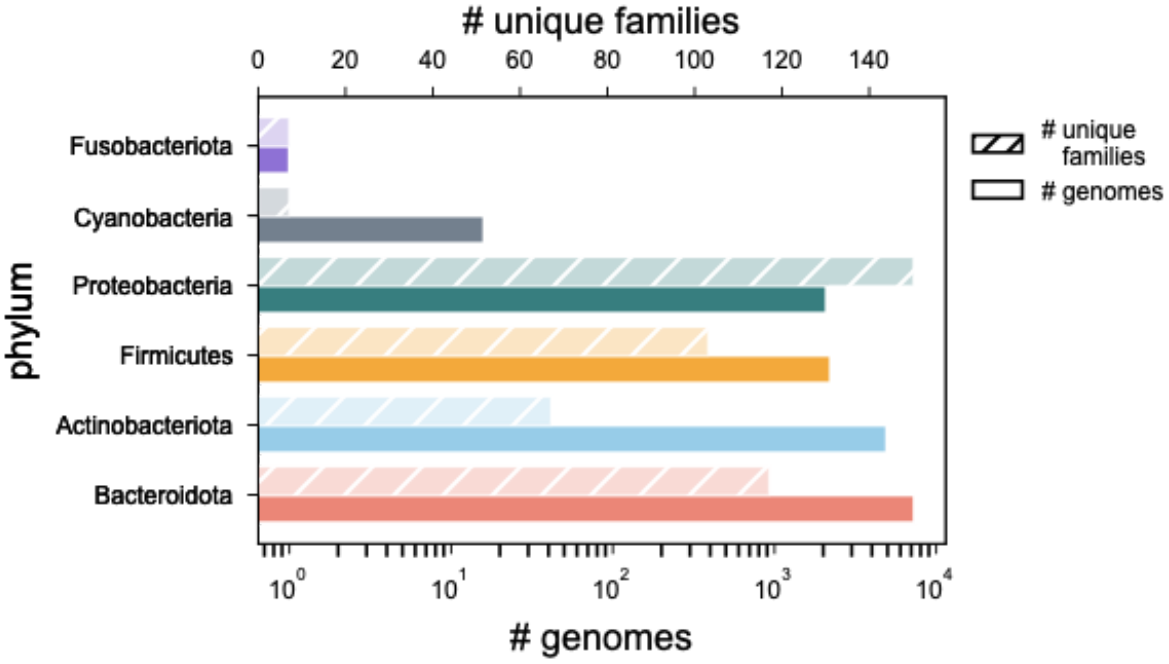


Figure 5. Proteobacteria comprise a minority of StrainAtlas genomes but contain a plurality of antiphage systems.

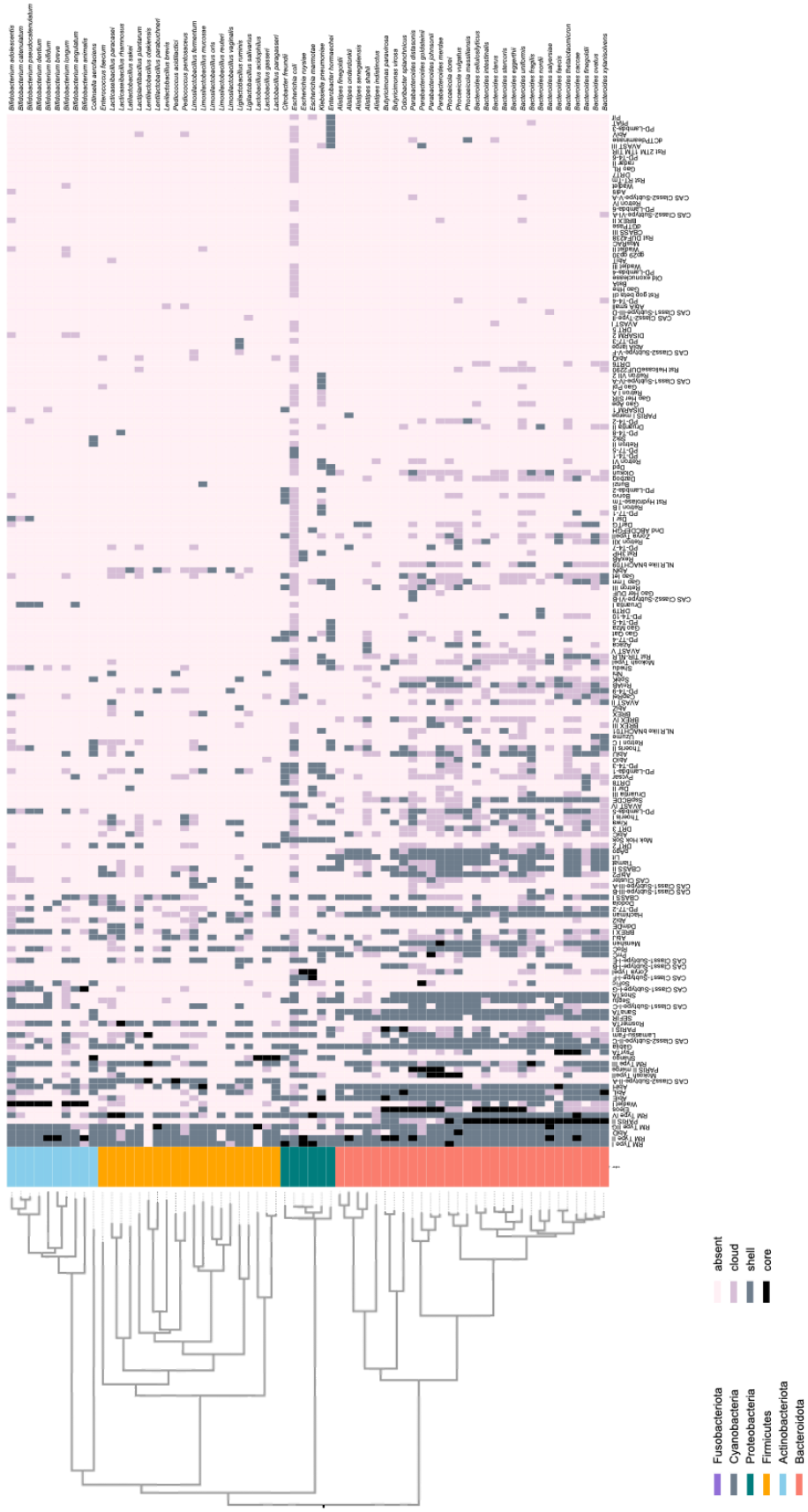

absent
cloud
shell
core
Fusobacteriota
Cyanobacteria
Proteobacteria
Firmicutes
Actinobacteriota
Bacteroidota

Figure 6. The majority of all detected antiphage families in StrainAtlas can be found in Proteobacteria (shown are StrainAtlas species with ≥10 genomes).

Most known antiphage defense systems remain rare in large species pangenomes

To examine the frequency of antiphage defense families within species, we consider each family in the context of a species pangenome, constructed using every strain of that species in StrainAtlas. Focussing first on the example of *E. coli*, which is well-represented in StrainAtlas with over 1000 genomes, we find that the vast majority of newly-discovered families detected in *E. coli* genomes belong to the cloud section of its pangenome (which we define as genes present in <5% of all genomes) (**Figure 7**). In other words, most of these families are exceedingly rare, even among such a sizable genome set. We next take one step back to include all unique families found in *E. coli*. By performing repeated (n=100) randomized resampling of genome sets of size 1 to all *E. coli* genomes, we observe that as the pangenome size increases, most families continue to be present only in the cloud section, and that this levels out at ~75% of families (**Figure 8**). This leveling out indicates that the rarity of these families in StrainAtlas likely reflects their underlying overall rarity in the *E. coli* species rather than our mistakenly-assigning non-rare families as rare due to our undersampling. This then suggests that additional currently-unknown rare families may be discovered with increased genome sampling. Repeating this analysis for all species with ≥100 genomes, we find that this generally holds true across phyla in StrainAtlas. Together, these findings suggest that continuing deep intra-species sampling is crucial to capture the true diversity of that species' antiphage arsenal.

Antiphage System	Prevalence in pangenome
PD-Lambda-1	10.65%
PD-Lambda-2	0.21%
PD-Lambda-3	0.54%
PD-Lambda-4	2.09%
PD-Lambda-5	0.75%
PD-Lambda-6	0.43%
PD-T4-1	8.51%
PD-T4-10	0.43%
PD-T4-2	1.44%
PD-T4-3	13.38%
PD-T4-4	0.00%
PD-T4-5	1.18%
PD-T4-6	0.16%
PD-T4-7	0.05%
PD-T4-8	2.94%
PD-T4-9	0.16%
PD-T7-1	2.09%
PD-T7-2	3.64%
PD-T7-3	0.32%
PD-T7-4	2.14%
PD-T7-5	9.42%

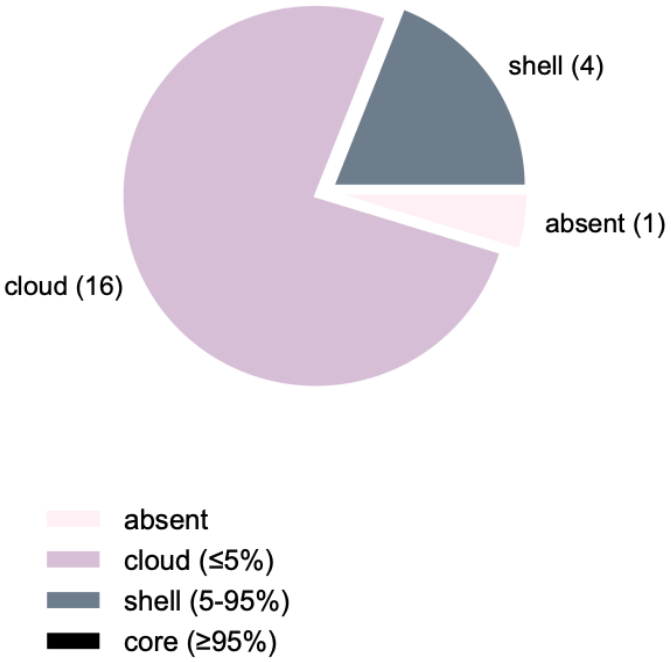
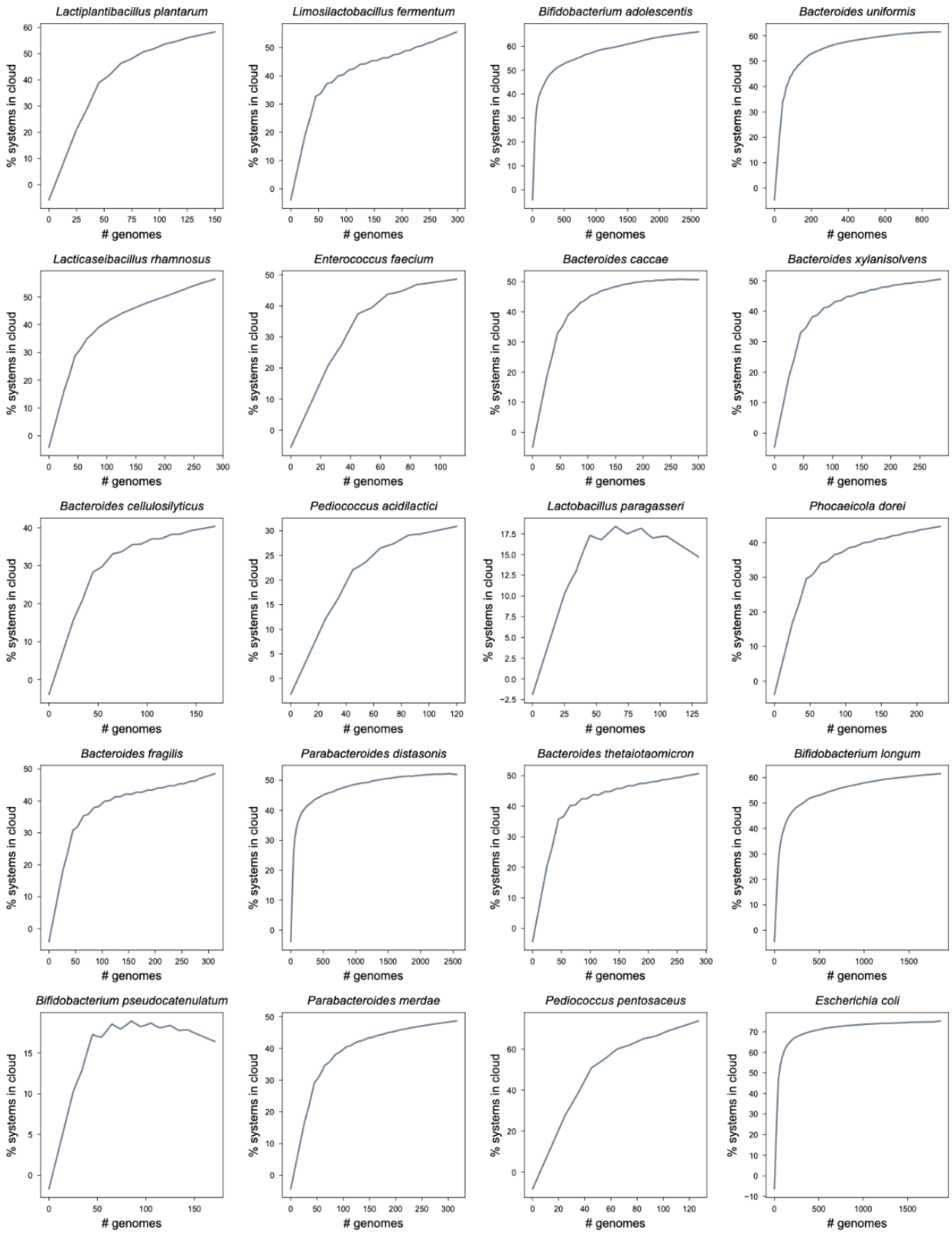


Figure 7. Newly-discovered antiphage systems from Vassallo et al. are enriched in the cloud section of the *E. coli* pangenome.


Lactiplantibacillus plantarum
Limosilactobacillus fermentum
Bifidobacterium adolescentis
Bacteroides uniformis
Lacticaseibacillus rhamnosus
Enterococcus faecium
Bacteroides caccae
Bacteroides xylanisolvens
Bacteroides cellulosilyticus
Pediococcus acidilactici
Lactobacillus paragasseri
Phocaeicola dorei
Bacteroides fragilis
Parabacteroides distasonis
Bacteroides thetaiotaomicron
Bifidobacterium longum
Bifidobacterium pseudocatenulatum
Parabacteroides merdae
Pediococcus pentosaceus
Escherichia coli
% systems in cloud
# genomes

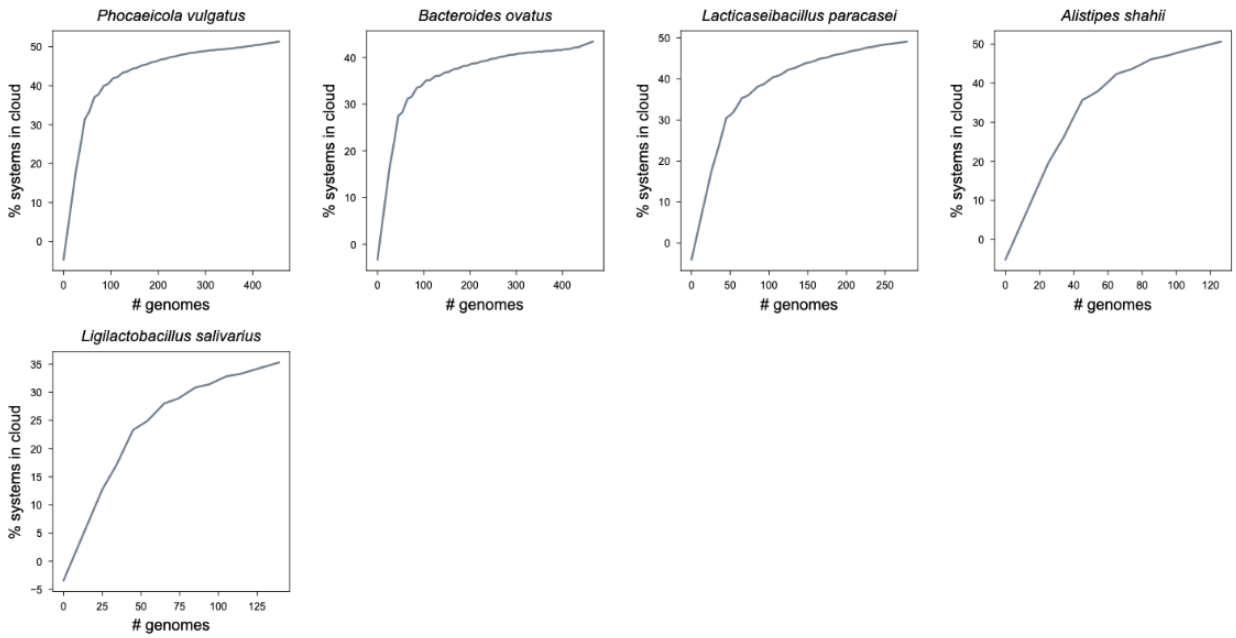


Figure 8. As we increase the genome set size, most antiphage systems remain in the cloud section of the pangenome, indicating their rarity in StrainAtlas reflects true rarity in the species. Shown are all StainAtlas species with ≥100 genomes.

Most species appear to fall far short of E. coli's defense repertoire

We next compare pangenomes cross-species to understand the relative size of the phage defense capacity of each species as a whole. We see that pangenome-wide unique family count correlates positively with species average genome size (Pearson r=0.39, p=4.67e-08), but that *E. coli* is again an outlier, with >50% more unique defense families in its pangenome versus the next species with a comparable genome size (**Figure 9a**). Indeed, its pangenome-wide defense arsenal far outstrips every other species, despite having only a middle-of-the-pack genome size (5.06Mbp) and per-genome average family count (8.97). To mitigate potential bias from the relatively high representation of *E. coli* in StrainAtlas, we repeat this analysis including only species with ≥100 genomes and randomly downsampling to exactly 100 genomes (**Figure 9b**), and also visualize pangenome-wide defense unique family count versus total pangenome size (**Figure 9c**). In both cases, *E. coli* remains a clear outlier species, with a much more expansive antiphage defense repertoire than we would expect for a typical species of its size. In turn, this reflects the fact that the relatively smaller

apparent size of defense repertoires of non-*E. coli* species may stem from lack of investigation rather than their true modest capacity.

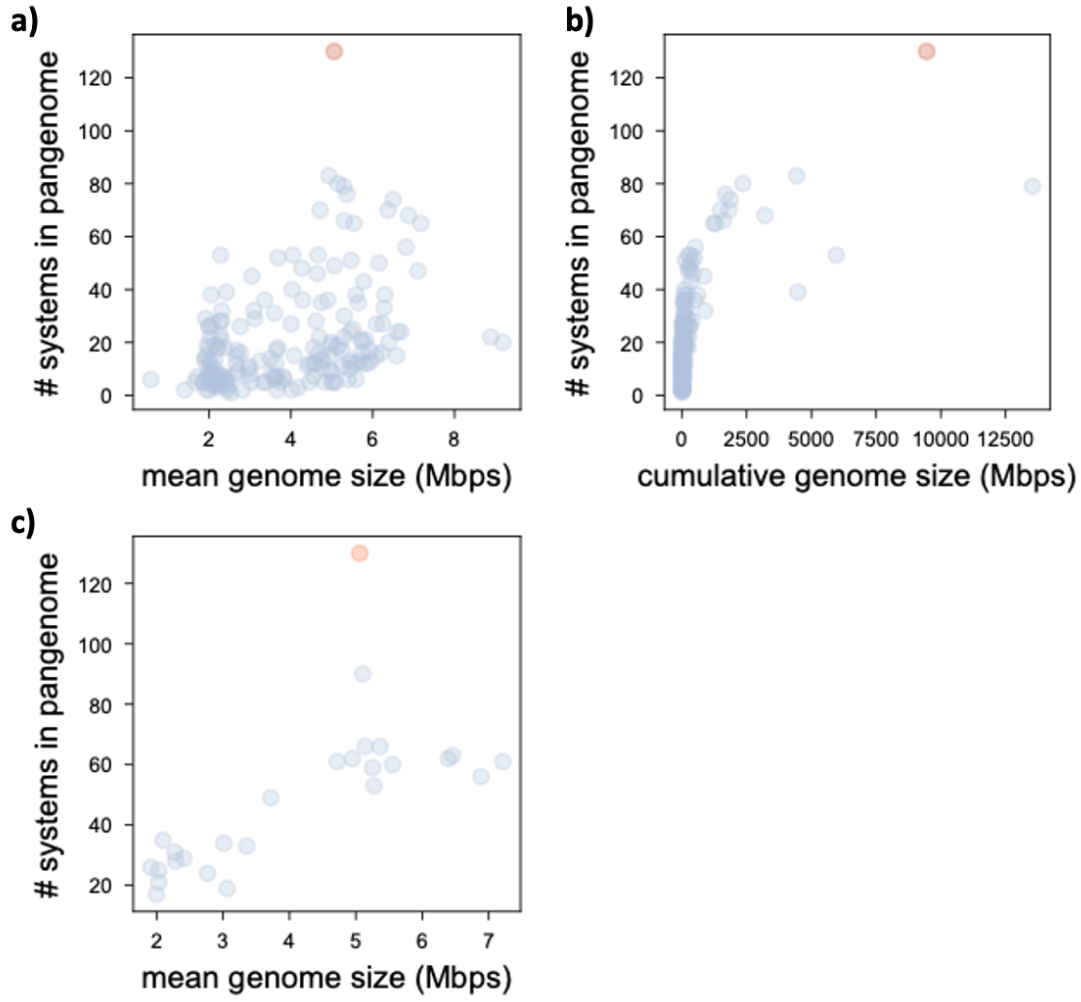


Figure 9. a) Total unique antiphage family count in a species positively correlates with species average genome size, but *E. coli* (in red) is an outlier with an outsized number of families. b) Taking into account varying species representation in StrainAtlas by plotting against cumulative sum of genome sizes within species, *E. coli* remains an outlier. c) Repeating a) but including only species with ≥100 genomes and randomly subsampling to 100 genomes, again to mitigate varying species representation, *E. coli* remains an outlier.

There is likely additional undiscovered diversity in E. coli's own antiphage defense repertoire

We now pivot to focus on *E. coli* itself, which as mentioned, is perhaps the best studied of gut (and all) bacteria. In spite of this, we hypothesize that there is much value in continuing to explore *E. coli*'s viral defense and that much like other non-model bacteria, the key lies in in-depth sampling of intra-species diversity. A recent paper from colleagues at MIT takes this approach and succeeds in identifying 21 previously-unknown defense system families by applying an experimental selection scheme on 71 diverse strains of *E. coli*. Using their data, we calculate the portion of known and novel families at genome set sizes downsampled from 1 to 71, and fit a logistic curve to the novel family counts using the scipy.optimize.curve_fit() function (RMSE=0.13, R^2=0.9994) (**Figure 10a**). We use the fit equation to extrapolate the total number of novel families expected in an *E. coli* genome set size equal to that of StrainAtlas' (50 families in n=1869 genomes) (**Figure 10b**). This reflects the uncharted territory of *E. coli* phage defense that could be discovered by applying Vassallo et al.'s technique to StrainAtlas. In fact, this could very well be a conservative estimate as it is beholden to their technique's biases and limitations.

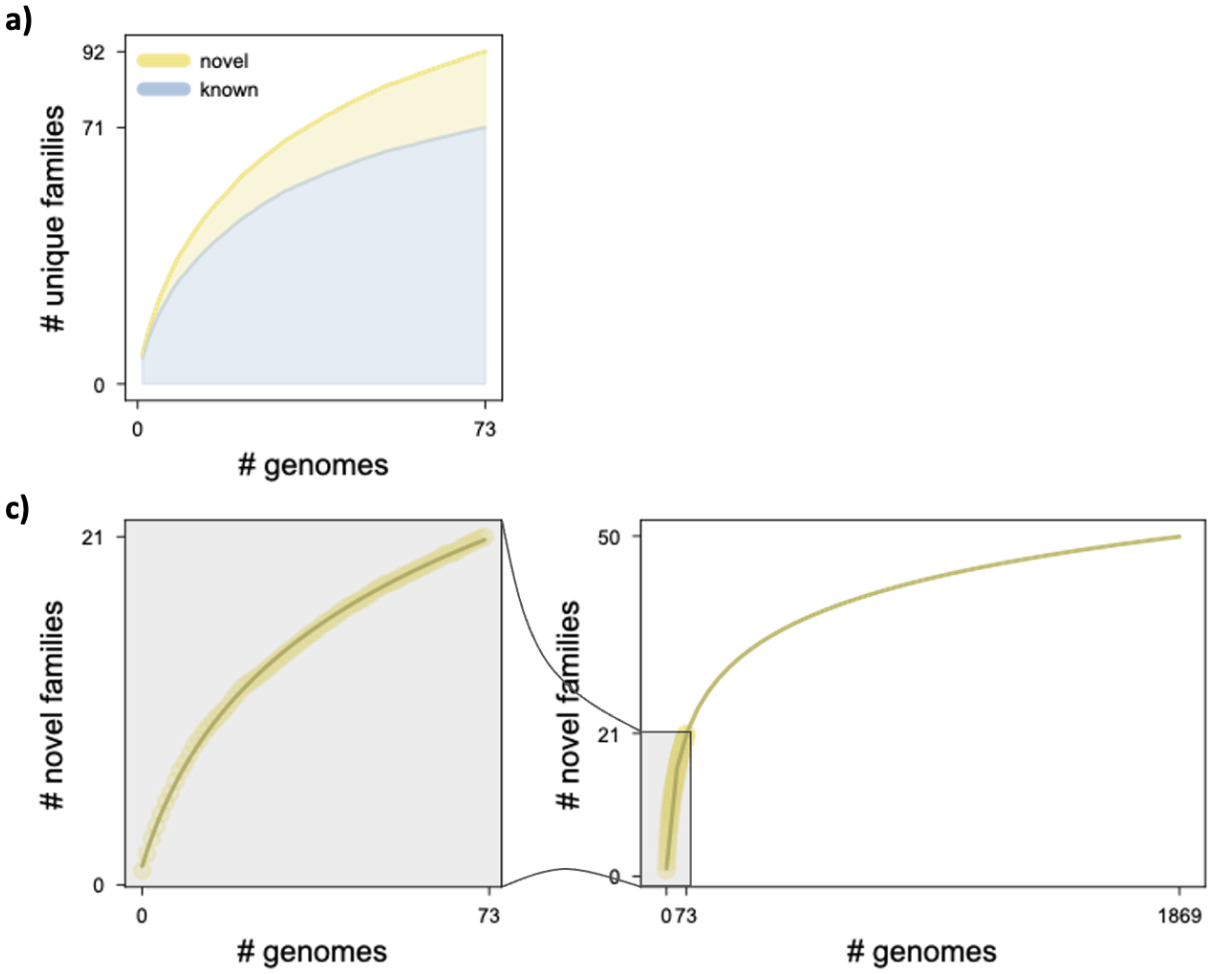
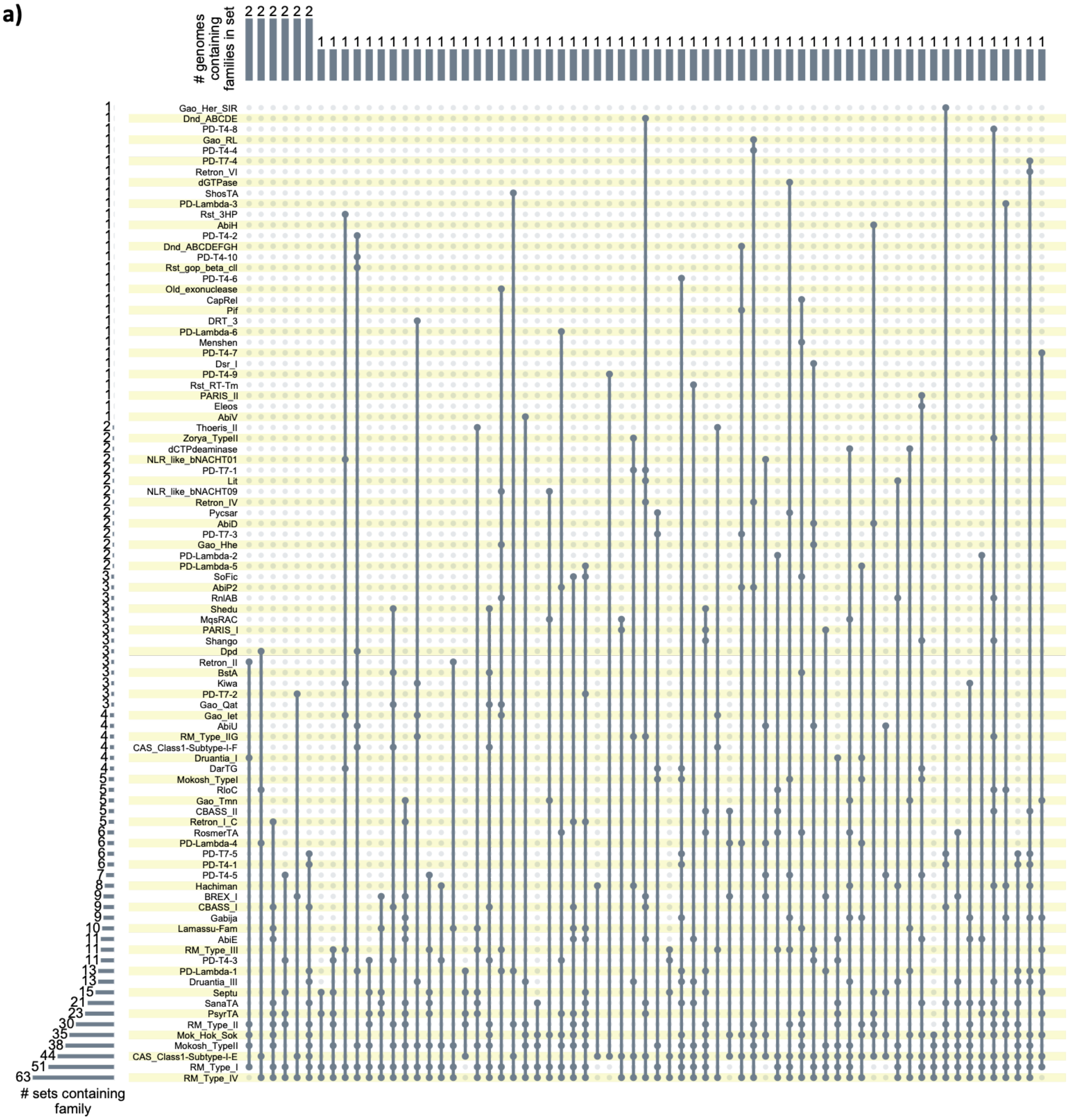


Figure 10. a) As the number of genomes increases, so too do the number of both known and novel antiphage systems in the *E. coli* genomes from Vassallo et al., without appearing to reach saturation. b) Using the occurrence patterns observed in the Vassallo et al. *E. coli*, we extrapolate to estimate the number of novel antiphage systems to be discovered within the StrainAtlas *E. coli* collection.

a)



genomes containing families in set
sets containing family
Gao_Her_SIR
Dnd_ABCDE
PD-T4-8
Gao_RL
PD-T4-4
PD-T7-4
Retron_VI
dGTPase
ShosTA
PD-Lambda-3
Rst_3HP
AbiH
PD-T4-2
Dnd_ABCDEFGH
PD-T4-10
Rst_gop_beta_cll
PD-T4-6
Old_exonuclease
CapRel
Pif
DRT_3
PD-Lambda-6
Menshen
PD-T4-7
Dsr_I
PD-T4-9
Rst_RT-Tm
PARIS_II
Eleos
AbiV
Thoeris_II
Zorya_TypeII
dCTPdeaminase
NLR_like_bNACHT01
PD-T7-1
Lit
NLR_like_bNACHT09
Retron_IV
Pycsar
AbiD
PD-T7-3
Gao_Hhe
PD-Lambda-2
PD-Lambda-5
SoFic
AbiP2
RnlAB
Shedu
MqsRAC
PARIS_I
Shango
Dpd
Retron_II
BstA
Kiwa
PD-T7-2
Gao_Qat
Gao_Iet
AbiU
RM_Type_IIG
CAS_Class1-Subtype-I-F
Druantia_I
DarTG
Mokosh_TypeI
RloC
Gao_Tmn
CBASS_II
Retron_I_C
RosmerTA
PD-Lambda-4
PD-T7-5
PD-T4-1
PD-T4-5
Hachiman
BREX_I
CBASS_I
Gabija
Lamassu-Fam
AbiE
RM_Type_III
PD-T4-3
PD-Lambda-1
Druantia_III
Septu
SanaTA
PsyrTA
RM_Type_II
Mok_Hok_Sok
Mokosh_TypeII
CAS_Class1-Subtype-I-E
RM_Type_I
RM_Type_IV

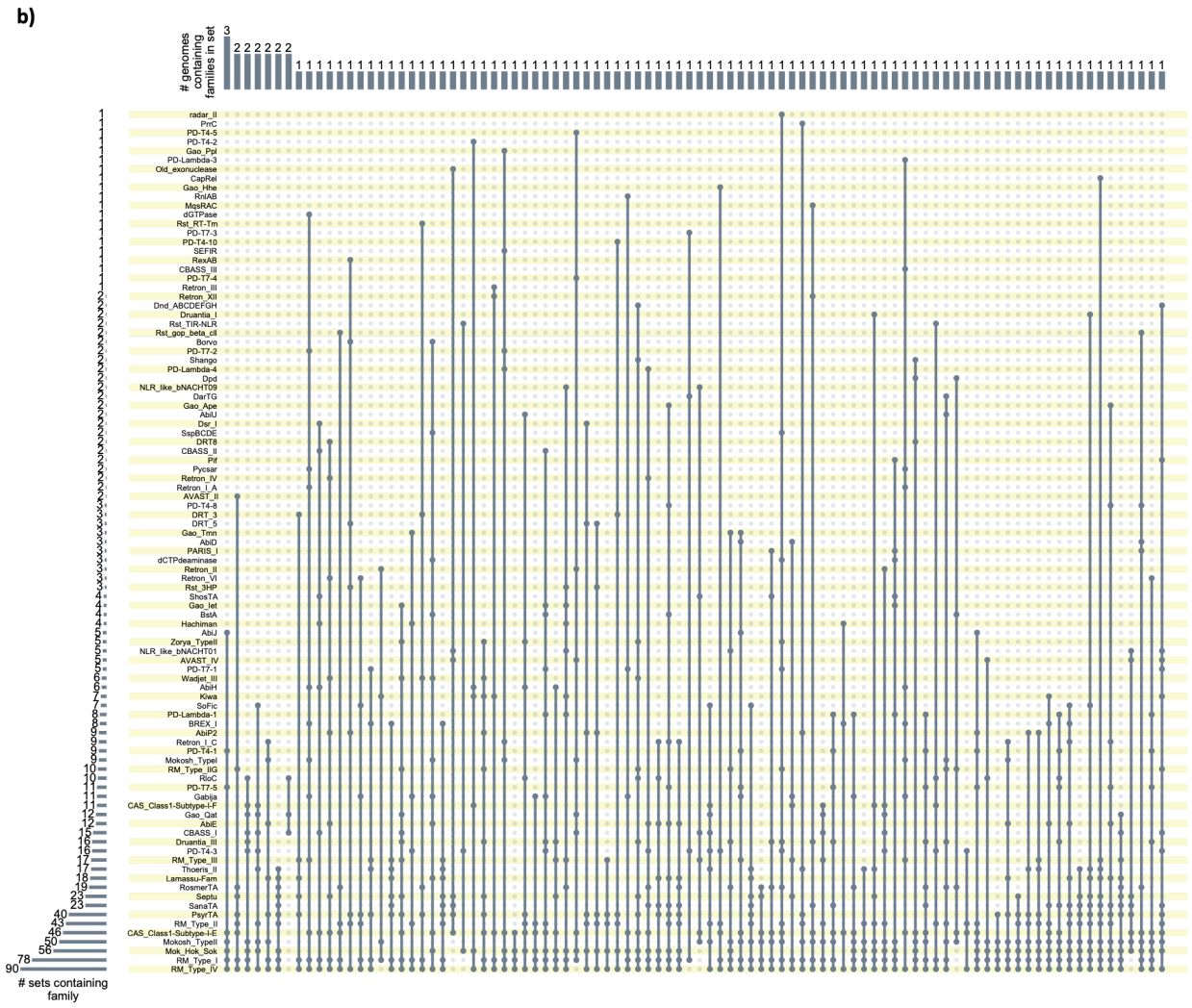


Figure 11. a) This upset plot depicts every unique combination of families found in the Vassallo et al. *E. coli* (vertical connected dots) and the number of genomes containing that exact combination (top histogram), showing wide dispersion of families across all genomes. b) The same plot for a random subsample of 100 StrainAtlas *E. coli* genomes.

We further probe the diversity within this dataset by calculating every set of unique combinations of defense families and for each set, tabulating the genomes that contain that exact combination (**Figure 11a**). The result demonstrates that families are spread widely and evenly throughout all genomes, rather than there being a concentration within a few genomes or a partitioning of certain combinations of families within groups of genomes. This can be seen from the high number of family sets (n=67), the modest average size of family sets (8.30 systems/set), and the very small genome counts

corresponding to each set (mode=1 genome/set). Repeating this analysis for a comparable number of randomly-subsampled StrainAtlas *E. coli* genomes (n=100), we observe similar results (number of family sets=92; mean family set size=8.62; mode of genome counts=1 genome/set) (**Figure 11b**). Once again, this reinforces the necessity for extensive intra-species sampling, as no arbitrary sparse downsampling of either of these *E. coli* datasets would suffice to capture the majority of the defense arsenal contained in the whole dataset, much less in the species as a whole.

Discussion

In this chapter we approach the question of bacterial strain-level diversity from one specific angle, focussing on strain-to-strain variability in viral defense across many gut species. With StrainAtlas we are able to show that the relatively modest defense arsenal equipped to the average gut bacterium usually belies the much larger arms selection available to its parent species. This global view of antiviral defense in the gut does away with the notion that any one or handful of strains is a sufficient representation for a species. Instead, researchers should move toward larger-scale systematic intra-species interrogations, now made much more accessible with our biobank.

Our survey of viral defense also points to the promising horizon of defense mechanism discovery, both in current StrainAtlas isolates and in future broad culture-and-sequencing efforts. We show that for model and non-model bacteria alike, even StrainAtlas' extensive sampling depths do not appear to exhaust diversity at the species level, and that many new and rare defense families are likely yet to be discovered. *E. coli* showing up again and again as an outlier with an apparently huge pangenome-level defense repertoire serves to cast doubt on the sufficiency of repertoire detection in other species. In effect, its own richness highlights what is missing elsewhere – what is unknown or cannot be found with current computational methods.

Here we acknowledge the limitations of our data and method. Our data, derived from a culture-based procedure, is naturally biased by which bacteria are culturable in a lab

setting, and indeed in the theme of strain-level variation, which strains out of these are most easily culturable. Our computational method relies on an existing HMM database and thus can only recognize what is similar enough to systems already studied. Any wholly unique unstudied system would be missed entirely here, though our goal is to aim a light in their direction with our current results.

We hope that our data and results will empower future research toward expansion of our knowledge of phage defense, not only through discovery of new systems, but also extension of study into their synergy, sharedness with components of eukaryote immunity, and more.

References

Bernheim, Aude, et al. “Prokaryotic Viperins Produce Diverse Antiviral Molecules.” *Nature*, vol. 589, no. 7840, Sept. 2020, pp. 120–24.

Blackwell, Grace A., et al. “Exploring Bacterial Diversity via a Curated and Searchable Snapshot of Archived DNA Sequences.” *PLoS Biology*, vol. 19, no. 11, Nov. 2021, p. e3001421.

Chambers, Jenni, and Thomas Arron Illingworth. “Bacterial Interactions Affecting Chemotherapy Effectiveness.” *McGill Science Undergraduate Research Journal*, vol. 18, no. 1, Apr. 2023, pp. B15–18.

Costea, Paul I., et al. “Enterotypes in the Landscape of Gut Microbial Community Composition.” *Nature Microbiology*, vol. 3, no. 1, Jan. 2018, pp. 8–16.

Doron, Shany, et al. “Systematic Discovery of Antiphage Defense Systems in the Microbial Pangenome.” *Science*, vol. 359, no. 6379, Mar. 2018, https://doi.org/10.1126/science.aar4120.

Gao, Linyi, et al. “Diverse Enzymatic Activities Mediate Antiviral Immunity in Prokaryotes.” *Science*, vol. 369, no. 6507, Aug. 2020, pp. 1077–84.

Hyatt, Doug, et al. “Prodigal: Prokaryotic Gene Recognition and Translation Initiation Site Identification.” *BMC Bioinformatics*, vol. 11, no. 1, Mar. 2010, pp. 1–11.

Jolley, Keith A., et al. “Ribosomal Multilocus Sequence Typing: Universal Characterization of Bacteria from Domain to Strain.” *Microbiology*, vol. 158, no. Pt 4, Apr. 2012, pp. 1005–15.

Millman, Adi, et al. “An Expanded Arsenal of Immune Systems That Protect Bacteria from Phages.” *Cell Host & Microbe*, vol. 30, no. 11, Nov. 2022, pp. 1556–69.e5.

Park, Sang-Cheol, et al. “Large-Scale Genomics Reveals the Genetic Characteristics of Seven Species and Importance of Phylogenetic Distance for Estimating Pan-Genome Size.” *Frontiers in Microbiology*, vol. 10, Apr. 2019,

https://doi.org/10.3389/fmicb.2019.00834.
Payne, Leighton J., et al. “Identification and Classification of Antiviral Defence Systems in Bacteria and Archaea with PADLOC Reveals New System Types.” *Nucleic Acids Research*, vol. 49, no. 19, Nov. 2021, pp. 10868–78.
Poirel, Laurent, et al. “Antimicrobial Resistance in.” *Microbiology Spectrum*, vol. 6, no. 4, July 2018, https://doi.org/10.1128/microbiolspec.ARBA-0026-2017.
Tesson, Florian, et al. “Systematic and Quantitative View of the Antiviral Arsenal of Prokaryotes.” *Nature Communications*, vol. 13, no. 1, May 2022, pp. 1–10.
Tettelin, Hervé, et al. “Genome Analysis of Multiple Pathogenic Isolates of Streptococcus Agalactiae: Implications for the Microbial ‘Pan-Genome.’” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 39, Sept. 2005, pp. 13950–55.
Wick, Ryan R., et al. “Unicycler: Resolving Bacterial Genome Assemblies from Short and Long Sequencing Reads.” *PLoS Computational Biology*, vol. 13, no. 6, June 2017, p. e1005595.
Wolf, Yuri I., et al. “Updated Clusters of Orthologous Genes for Archaea: A Complex Ancestor of the Archaea and the Byways of Horizontal Gene Transfer.” *Biology Direct*, vol. 7, Dec. 2012, p. 46.
Yang, Chao, et al. “Fecal IgA Levels Are Determined by Strain-Level Differences in Bacteroides Ovatus and Are Modifiable by Gut Microbiota Manipulation.” *Cell Host & Microbe*, vol. 27, no. 3, Mar. 2020, pp. 467–75.e6.

Chapter 3: Longitudinal immunosequencing in healthy people reveals persistent T cell receptors rich in highly public receptors

This chapter is a collaboration between me, Nathaniel Chu, and scientists at Adaptive Biotechnologies. The Adaptive scientists performed all sample collection and sequencing, Nathaniel and I performed all analyses, Nathaniel performed visualization, and I wrote the manuscript. This chapter is adapted from the published paper in BMC Immunology (2019).

Abstract

Background

The adaptive immune system maintains a diversity of T cells capable of recognizing a broad array of antigens. Each T cell's specificity for antigens is determined by its T cell receptors (TCRs), which together across all T cells form a repertoire of millions of unique receptors in each individual. Although many studies have examined how TCR repertoires change in response to disease or drugs, few have explored the temporal dynamics of the TCR repertoire in healthy individuals.

Results

Here we report immunosequencing of TCR β chains (TCRβ) from the blood of three healthy individuals at eight time points over one year. TCRβ repertoires of all peripheral-blood T cells and sorted memory T cells clustered clearly by individual, systematically demonstrating that TCRβ repertoires are specific to individuals across time. This individuality was absent from TCRβs from naive T cells, suggesting that the differences resulted from an individual's antigen exposure history, not genetic background. Many characteristics of the TCRβ repertoire (e.g., diversity, clonality) were stable across time, although we found evidence of T cell expansion dynamics even within healthy individuals. We further identified a subset of "persistent" TCRβs present across all time points. These receptors were rich in clonal and highly public receptors and may play a key role in immune system maintenance.

Conclusions

Our results highlight the importance of longitudinal sampling of the immune system, providing a much-needed baseline for TCRβ dynamics in healthy individuals. Such a baseline will improve interpretation of changes in the TCRβ repertoire during disease or treatment.

Background

T cells play a vital role in cell-mediated immunity, one branch of the adaptive immune response against foreign and self-antigens. Upon recognizing an antigen from an antigen-presenting cell, naive T cells activate and proliferate rapidly. This process stimulates an effector response to the immediate challenge, followed by generation of memory T cells, which form a lasting cohort capable of mounting more-efficient responses against subsequent challenges by the same antigen.

The key to the flexibility and specificity of T cell responses lies in the cells' remarkable capacity to diversify their T cell receptor (TCR) sequences, which determine the antigens those cells will recognize. Most T cells display TCRs made up of two chains: an α and a β chain. Sequence diversity in these chains arises during T cell development, through recombination of three sets of gene segments: the variable (V), diversity (D), and joining (J) segments [1]. Random insertions and deletions at each genetic junction introduce still more diversity, resulting in a theoretical repertoire of 10^{15} unique receptors in humans [2]. Selective pressures during and after T cell development, as well as constraints on the number of T cells maintained by the body, limit this diversity to an observed 10^{7} (approximately) unique receptors per individual [2-5].

This TCR repertoire forms the foundation of the adaptive immune response, which dynamically responds to disease. Each immune challenge prompts expansions and contractions of different T cell populations, and new T cells are continually generated. Substantial research interest has focused on these dynamics in the context of immune system perturbations, including in cancer [6-9], infection [10,11], autoimmune disorders [12,13], and therapeutic trials [8,14,15]. Observing changes in TCR populations not only uncovers cellular mechanisms driving disease, but can inform development of new diagnostics, biomarkers, and therapeutics involving T cells.

Less research has explored TCR dynamics in healthy individuals. Previous studies found that some TCRs remain present in individuals over decades [16,17], but these

long-term studies may not directly relate to shorter-term events, such as diseases or treatments. Interpreting TCR dynamics when the immune system is challenged would be more straightforward if we had a clear picture of TCR dynamics in healthy individuals.

To help develop this picture, we report immunosequencing of peripheral TCR β chain (TCRβ) repertoires of three individuals at eight time points over one year. We focused on the TCR β chain because, unlike the α chain, only one β chain can be expressed on each T cell [18]**,** the β chain contains greater sequence diversity [19], and it more frequently interacts with presented antigens during recognition [20]. These factors suggest that TCRβ sequences should be sufficient to track individual T cells and their clones. Our analysis revealed overall individuality and temporal stability of the TCRβ pool. We also uncovered a set of temporally persistent TCRβs, which were more abundant, and shared across more people, than transitory TCRβs.

Results

T cell receptor repertoires show individuality and stability through time

To characterize the dynamics of T cell receptors in healthy individuals, we deeply sequenced the TCRβ locus of all T cells from peripheral-blood mononuclear cells (PBMCs) isolated from three healthy adults (for schematic of experimental design, see **Figure 1a**). We sampled each individual at eight time points over one year (**Figure 1a**). For three intermediate time points, we also sequenced flow-sorted naive and memory T cells from PBMCs (see Methods). Our deep sequencing effort generated ~21 million (+/- 6 million SD) sequencing reads and -250,000 (+/- 100,000 SD) unique, productive TCRβs – which we defined as a unique combination of a V segment, CDR3 amino acid sequence, and J segment [21] – per sample. These values and other summary statistics per sample appear in **Table S1.** Most TCRβs had abundances near 10^{-6} (**Figure S1**), and rarefaction curves indicate that all samples were well saturated (**Figure S2**). This saturation indicates that our sequencing captured the full diversity of TCRβs in our

samples, although our blood samples cannot capture the full diversity of the TCRβ repertoire (see **Discussion**).

We first examined whether previously observed differences among individuals were stable through time [7,22]. Looking at shared TCRβs (Jaccard index) among samples, we indeed found that samples of PBMCs or memory T cells taken from the same individual shared more TCRβs than samples taken from different individuals (**Figure 1b**), and this pattern was consistent over one year. In adults, memory T cells are thought to make up 60-90% of circulating T cells [23,24], which aligns with the agreement between these two T cell sample types. In contrast, TCRβs from naive T cells did not cluster cohesively by individual (**Figure 1b**)**.** As naive T cells have not yet recognized a corresponding antigen, this lack of cohesion might suggest one of two possibilities: (1) that before antigen recognition and proliferation, TCRβ repertoires are not specific to individuals or (2) the naive T repertoire is simply too diverse or too dynamic for individuality to manifest. We thus conclude that at the depth of sequencing and sampling of this study, individuality results from an individual's unique antigen exposure and T cell activation history, which shape memory and total T cell repertoires.

We next examined patterns across samples from the same individual to understand TCR dynamics in healthy individuals. We observed only a minority of TCRβs shared among samples from month to month; indeed, samples of PBMCs at different months from the same individual typically shared only 11% of TCRβs *(+/-* 3.6% SD, range 5-18%) (**Figure 1b**).

Two factors likely played a role in the observed turnover of TCRβ repertoires: (1) changes in TCRβ abundances in the blood across time and (2) inherent undersampling of such a diverse system (see **Discussion**). Surveying peripheral blood immune repertoires undersamples at multiple points, including blood drawing, nucleic acid extraction, library construction, and sequencing. The resulting undersampling likely explained much of the low overlap of TCRβs among samples but simultaneously highlighted the significance of TCRβs shared across time points. To verify that patterns we observed were not artifacts of undersampling, we also analyzed a subset of

high-abundance TCRβs (those ranked in the top 1% by abundance, see **Methods**, **Additional File 1**), which are less likely to be affected. In these TCRβs, we observed typical sharing of 63% *(+/-* 13.8% SD, range 35-88%) of TCRβs in PBMC samples across time (**Figure S3a**). PBMC and memory T cell samples (but not naive T cell samples) still clearly clustered by individual when only these TCRβs were considered (**Figure S3a**).

The frequencies of high-abundance TCRβs from each individual were largely consistent over time (**Figure 1c).** We found that abundances of the same TCRs correlated within individuals over the span of a month (**Figure 1d, S3b)** and a year (**Figure 1e, S3c**). This correlation was particularly strong for abundant TCRβs **(Figure S3b-c**) whereas rare TCRβs varied more. This correlation held true in naive and memory T cell subpopulations, sampled across a month (**Figure 1f-g**). In contrast, correlation was much weaker among abundances of TCRβs shared across individuals (**Figure 1h, S3d**), again highlighting the individuality of each repertoire. We found that the proportion of shared TCRβs (Jaccard index) tended to decrease with longer time intervals passed between samples, although with a notable reversion in Individual 02 (**Figure S4**). We observed stable diversity **(Figure 1i, S3e**), clonality **(Figure 1j, S3f),** and V and J usage (**Figure S5, S6; Table S2, S3**) within individuals over time.

In the absence of experimental intervention, we observed complex clonal dynamics in many TCRβs, including cohorts of TCRβs with closely correlated expansion patterns (**Figure S7**). To avoid artifacts from undersampling, we looked for such cohorts of correlating receptors only in high-abundance TCRβs (see **Methods).** In all individuals, many high-abundance TCRβs appeared together only at a single time point. We also found cohorts of high-abundance TCRβs that correlated across time points (**Figure S7**). Some of these cohorts included TCRβs that fell across a range of abundances (**Figure S7a-b**), while other cohorts were made up of TCRβs with nearly identical abundances (**Figure S7c**). Correlating TCRβs were not obviously sequencing artifacts (**Table S4, Methods**). These cohorts of closely correlated TCRβs indicate that even in healthy individuals whose overall TCR repertoire appears stable, underlying dynamics remain.

Taken together, these results revealed a diverse system, which nevertheless displayed consistent, unifying features differentiating individuals, plus longitudinal dynamics that suggested continual immune processes.

A persistent TCRβ repertoire contains elevated proportions of clonal, highly public TCRβs

During our analysis, we discovered a subset of TCRβs that was present across all eight PBMC samples from a single individual, a subset we called "persistent" TCRβs (**Figure 2a**). While approximately 90% of unique TCRβs observed over all of an individual's PBMC samples occurred in only one sample, 0.3-0.8% of TCRβs occurred at all eight time points (**Figure 2a**). When considering individual samples, this pattern translated to 1-5% of TCRβs observed in each sample were persistent receptors (**Table S5**). When we considered only high-abundance TCRβs, the frequency of persistent TCRβ increased substantially (**Figure S8a**).

We hypothesized that these persistent TCRβs might be selected for and maintained by the immune system, perhaps to respond to continual antigen exposures or other chronic immunological needs.

In our data, we found multiple signatures of immunological selection acting on persistent TCRβs. The members of this persistent subset tended to have a higher mean abundance than TCRβs observed at fewer time points (**Figure 2b, Table S6**). We also observed that the number of unique nucleotide sequences encoding each TCRβ's CDR3 amino acid sequence was generally higher for persistent TCRβs (**Figure 2c, Table S7**). This pattern of greater nucleotide redundancy varied across individuals and region of the CDR3 sequence (**Figure S9a**), but TCRβs with the highest nucleotide redundancy were reliably persistent (**Figure S9b**). Furthermore, we discovered that TCRβs occurring at more time points, including persistent TCRβs, shared larger proportions of TCRβs also associated with memory T cells (**Figure 2d**). Remarkably, 98% of persistent TCRβs also occurred in memory T cells, suggesting that almost all persistent T cell clones had previously encountered and responded to their corresponding antigens. We found a similar pattern in naive T cells, although the overall

overlap was lower (50%), indicating that persistent TCRβs were also enriched in the naive compartment **(Figure 2e**). Persistent TCRβs did not show altered CDR3 lengths or VJ usage (**Figure S10-S12**). Like alpha diversity and clonality, the cumulative abundance of TCRβs present in different numbers of samples appeared stable over time and specific to individuals (**Figure 2f**). Surprisingly, although persistent TCRβs constituted less than 1% of all unique TCRβs, they accounted for 10- 35% of the total abundance of TCRβs in any given sample (**Figure 2f**), further evidence that these T cell clones had expanded. We observed similar patterns when analyzing only high-abundance TCRβs (**Figure S8**).

Taken together, these characteristics – persistence across time, higher abundance, redundant nucleotide sequences, and overlap with memory T cells – suggest immunological selection for persistent TCRβs. We therefore investigated whether persistent TCRβs coexisted with TCRβs having very similar amino acid sequences. Previous studies have suggested that TCRβs with similar sequences likely respond to the same or similar antigens, and such coexistence may be evidence of immunological selection [25,26].

To explore this idea, we applied a network clustering algorithm based on Levenshtein edit distance between TCRβ CDR3 amino acid sequences in our data [25-27]. We represented antigen-specificity as a network graph of unique TCRβs, in which each edge connected a pair of TCRβs with putative shared specificity. We found that TCRβs having few edges-and thus few other TCRβs with putative shared antigen specificity-tended to occur in only one sample, while TCRβs with more edges included a higher frequency of TCRβs occurring in more than one sample (**Figure 3a**, $p < 10^{-5}$ for all three individuals by a nonparametric permutation test). This pattern indicates that TCRβs occurring with other, similar TCRβs were more often maintained across time in the peripheral immune system.

We next examined the association between persistent TCRβs-those shared across time points – and "public" TCRβs – those shared across people. Public TCRs show many of the same signatures of immunological selection as persistent TCRβs, including higher

abundance [28], overlap with memory T cells [28], and coexistence with TCRs with similar sequence similarity [25]. To identify public TCRβs, we compared our data with a similarly generated TCRβ dataset from a large cohort of 778 healthy individuals [21] (**Additional File 2**). We found that the most-shared (i.e., most-public) TCRβs from this large cohort had a larger proportion of persistent TCRβs from our three sampled individuals (**Figure 4a-b, Table S8**, $p < 10^{-5}$ for all three individuals by a nonparametric permutation test). Private TCRβs – those occurring in few individuals – most often occurred at only a single time point in our analyses. Interestingly, TCRβs that occurred at many but not all time points (i.e., 3-5 time points) were on average the most-shared (**Figure S14a**), but persistent TCRβs were specifically enriched in highly public TCRβs-here defined as those shared by over 70% of subjects in the large cohort (**Figure 4c, S14b**). The three most public TCRs (found in over 90% of the 778-individual cohort) were found to be in the persistent TCRβ repertoires of all three individuals and were diverse in structure (**Figure 4d**).

Public TCRs are thought to be products of genetic and biochemical biases in T cell receptor recombination [29,30] and also of convergent selection for TCRs that respond to frequently encountered antigens [21,32]. To better understand the effects of biases during TCRβ recombination on receptor persistence, we used IGoR to estimate the probability that each TCRβ was generated before immune selection [33]. Similar to previous studies [30], the probability that a given TCRβ was generated correlated closely with publicness (**Figure S15a**). In our time series data, TCRβs that occurred at multiple time points tended to have slightly higher generation probabilities than TCRβs only observed once (**Figure S15b),** but persistent TCRβs did not have higher generation probabilities than other receptors observed in more than one time point. In addition, more abundant TCRβs (both persistent and nonpersistent) did not have higher generation probabilities (**Figure S15c-d**). These results suggest that, unlike public receptors, persistent receptors and their abundances do not appear to result from biases in TCR recombination. The contradiction that public and persistent receptors are associated but only public TCRβs appear to be generated by recombination bias is possible because despite their association, these two TCRβ subsets are largely independent. Although the most public receptors are overwhelmingly persistent (**Figure**

4), they represent a tiny fraction of the persistent receptors in each individual. Thus, although these two subsets of the TCR repertoire-persistent and public-overlap and share many characteristics, they are also distinct, suggesting that they may play complementary roles in adaptive immunity.

Discussion

Our analyses revealed both fluctuation and stability in the TCRβ repertoire of healthy individuals, providing a baseline framework for interpreting changes in the TCR repertoire. We identified a number of consistent repertoire characteristics (e.g., diversity, clonality), which are known to be affected by immunizations, clinical interventions, and changes in health status [7,14,34]. These patterns differed among individuals across time, highlighting the role played by genetics [like human leukocyte antigen (HLA) type] and history of antigen exposure in shaping the TCR repertoire. We did not obtain HLA-type information from these three subjects, so the relative contributions of HLA type versus individual history remains unknown.

We further discovered a subset of persistent TCRβs that bore signs of immune selection. Persistent TCRβs tended to be more abundant than nonpersistent receptors, although this distinction is to a certain extent confounded by the fact that high-abundance receptors are also more likely to be detected in a given sample. Nevertheless, this circular logic does not detract from the immune system's maintenance of specific dominant TCRβs across time. We further found that persistent TCRβs had higher numbers of distinct nucleotide sequences encoding each TCRβ. TCR diversity is generated by somatic DNA recombination, so it is possible for the same TCR amino acid sequence to be generated from independent recombinations in different T cell clonal lineages. Thus, coexistence of multiple clonal lineages encoding the same TCRβ amino acid sequence may reflect selective pressures to maintain that TCRβ and its antigen specificity. Similarly, the presence of many TCRβs similar to persistent TCRβs – as identified by our network analysis – could also result from selection for receptors that recognize a set of related antigens [20,36]. Previous studies using network analyses also found that public TCRβs tend to occur with similar TCRβs

[25], further suggesting that both public and persistent TCRβs are key drivers of lasting immunity. In addition to using TCR sequencing to track TCRβs that proliferate in response to intervention, we propose that the three dimensions explored in this paper – similarity with other receptors, publicness across individuals, and persistence through time – represent useful strategies for identifying biologically important TCRβs.

The presence of near-ubiquitous (present in >90% of individuals in a cohort of 778 individuals) and persistent TCRβs led us to speculate that these TCRβs might be responding to a set of common antigens repeatedly encountered by healthy people. These antigens could be associated with self-antigens, chronic infections (e.g., Epstein-Barr virus), or possibly members of the human microbiota. In fact, the CDR3 sequence CASSPQETQYF has been previously been associated with the inflammatory skin disease psoriasis [37] and CASSLEETQYF has been implicated in responses to *Mycobacterium tuberculosis* [20] and cytomegalovirus [38].

In addition to persistent TCRβs, our analysis revealed many receptors with unstable, transient behavior. Many high-abundance TCRβs did not persist through time, with many occurring at only a single time point (**Figure 2b, S8a**). These TCRβs could well correspond to T cells that expanded during a temporary immune challenge but then did not persist in high abundance afterward. These dynamics might also reflect the migration of T cells to and from different tissues, which could manifest as fluctuating abundance in the blood. The presence of dynamically expanding or migrating TCRβs in apparently healthy individuals poses an important consideration for designing studies monitoring the immune system. Studies tracking TCR abundances in cross-sectional immune system sampling [7,14,34-36] may capture not only T cell clones responding to intervention, but also expanding clones inherent in the T cell dynamics of healthy individuals. Repeated sampling before and after intervention could minimize such false positives.

Current immunosequencing methods have limitations that should inform the interpretation of our results. Most important, given such a diverse system as the TCR repertoire, even large sequencing efforts like ours undersample. Although our

sequencing appeared to saturate our samples (**Figure S2**), additional bottlenecks during library preparation and, particularly, blood drawing limit our ability to capture full TCRβ diversity. Previous studies exhaustively sequenced multiple libraries from multiple blood samples, but even these estimates are considered a lower limit of TCRβ diversity [42]. This detection limit could confound our identification of persistent TCRβs. Many of the TCRβs that did not occur in all samples were undoubtedly present but too rare for our analysis to capture. Thus, identification of a persistent TCR repertoire was subject to an abundance cutoff, whereby we focused on TCRs that persisted above the detection limit of sampling. To check that our conclusions were not heavily altered by undersampling, we analyzed high-abundance TCRβs and found similar overall patterns, so we infer that our main conclusions are likely robust despite this experimental limitation. In addition, our study included data from only three female individuals ages 18-45. The immune system varies across sex [43] and age [44], and although the patterns we describe are clear, larger longitudinal studies on the immune repertoire with greater patient characterization (particularly HLA type) and representation (e.g., including men and a range of ages) will better define how these patterns apply across populations.

Conclusions

To better understand healthy immune system dynamics in humans, we profiled the TCRβ repertoires from three individuals over one year. We found a system characterized by both fluctuation and stability and further discovered a novel subset of the TCRβ repertoire that might play a key role in immunity. As immune profiling in clinical trials becomes more prevalent, we hope our results will provide much-needed context for interpreting immunosequencing data, as well as for informing future trial designs.

Methods

Study design

We sought to study baseline dynamics and characteristics of the TCRβ repertoire in healthy individuals across time. We sampled blood from three individuals from eight time points over one year. We kept our sample size small so that we could perform extremely deep immune repertoire profiling on each sample, a choice that should be taken into consideration when interpreting our results.

Sample collection

Three healthy adult female volunteers ages 18-45 provided blood samples over the course of one year, with samples taken on a starting date and 1, 2, 3, 5, 6, 7, and 12 months after that date (**Figure 1a**). We sequenced TCRβ chains from approximately 1 million PBMCs from each sample. From the samples at *5,* 6, and 7 months, we also sequenced TCRβ chains from sorted naive (CD3+, CD45RA+) and memory (CD3+, CD45RO+) T cells.

High-throughput TCR sequencing

We extracted genomic DNA from cell samples using a Qiagen DNeasy blood extraction kit (Qiagen, Gaithersburg, MD, USA). We sequenced CDR3 regions of rearranged TCRβ genes and defined these regions according to the international immunogenetics information system (IMGT) *[45].* We amplified and sequenced TCRβ CDR3 regions using previously described protocols [2,46]. Briefly, we applied a multiplexed PCR method, using a mixture of 60 forward primers specific to TCR Vβ gene segments plus 13 reverse primers specific to TCR Jβ gene segments. We sequenced 87 base-pair reads on an Illumina HiSeq System and processed raw sequence data to remove errors in the primary sequence of each read. To collapse the TCRβ data into unique sequences, we used a nearest-neighbor algorithm-merging closely related sequences-which removed PCR and sequencing errors. By sequencing genomic DNA and not RNA, our approach more accurately reflected T cell abundances but also captured both expressed and unexpressed T cell receptors [19].

Data analysis

In our analyses, we focused on TCRβs containing no stop codons and mapping successfully to a V gene and J gene (**Table S1**). Relative abundances of these "productive" TCRβ sequences, however, took into account the abundances of nonproductive TCRβ sequences, as these sequences were still part of the greater TCRβ pool. We defined a TCRβ as a unique combination of V gene, J gene, and CDR3 amino acid sequence. We examined nucleotide redundancy of each TCRβ by counting the number of T cell clones—a unique combination of V gene, J gene, and CDR3 nucleotide sequence— encoding each TCRβ. We defined TCRβs whose abundances ranked in the top 1% for each sample as high-abundance TCRβs, and we analyzed these TCRβs in parallel with the full TCRβ repertoire as a check for artifacts of undersampling (**Figure S5, S8**).

We calculated Spearman's and Pearson's correlation coefficients for TCRβ abundances across samples using the Python package SciPy, considering only TCRβs that were shared among samples. We calculated alpha diversity (Shannon estimate = $e^{(\text{Shannon entropy})}$) and clonality (1 - Pielou's evenness) using the Python package Scikit-bio *0.5.1.* We calculated Levenshtein distance using the Python package Python-Levenshtein 0.12.0 and analyzed the resulting network using the Python package NetworkX 1.9.1.

To look for TCRβs with similar temporal dynamics, we focused on TCRβs that occurred in the top 1% at least twice. These TCRβs likely represented T cell clones that had expanded. We then calculated Spearman's and Pearson's correlation coefficients for all high-abundance TCRβ pairs, filling in missing data with the median abundance of TCRβs from each sample. We used median abundance – instead of a pseudocount of 1 or half the minimum abundance detected – because the immense diversity of the TCRβ repertoire means that most detected TCRβs are likely similarly abundant as TCRβs that were not detected. We identified pairs of TCRβs that had high (>0.95) correlation. To identify cohorts of TCRβs that co-correlated, we represented TCRβs as nodes in a network, where nodes were connected by edges if the corresponding TCRβs were highly correlated. We then searched for the maximal network clique (a set of nodes where each node has an edge to all other nodes) using NetworkX. We visually inspected these TCRβ cohorts for evidence of sequencing error, which might have

resulted in a high-abundance TCRβ that closely correlated with many low-abundance TCRβs with similar sequences (**Table S4**). To test the significance of TCRβ cohort size, we performed the same analysis on 1000 shuffled datasets. Each shuffled dataset randomly permuted sample labels (i.e., the sampling date) for each TCRβ within each individual.

To test the significance of persistent TCRβ enrichment in (a) public receptors (**Figure 4**) and (b) TCRβs that occurred with many similar receptors (**Figure 3**), we analyzed 10,000 shuffled datasets. For these permutations, we randomly permuted the number of time points at which each TCRβ was observed and repeated the analysis.

We estimated the probability of generation of each TCRβ before immune selection using IGoR version 1.1.0 with the provided model parameters for the human TCRβ locus [33].

References

1. Chien Y, Gascoigne NRJ, Kavaler J, Lee NE, Davis MM. Somatic recombination in a murine T-cell receptor gene. Nature. 1984 May 24;309(5966):322-6.
2. Robins HS, Campregher PV, Srivastava SK, Wacher A, Turtle CJ, Kahsai 0, Riddell SR, Warren EH, Carlson CS. Comprehensive assessment of T-cell receptor β-chain diversity in αβ T cells. Blood. 2009 Nov 5;114(19):4099-107.
3. Arstila TP, Casrouge A, Baron V, Even J, Kanellopoulos J, Kourilsky P. A direct estimate of the human alphabeta T cell receptor diversity. Science. 1999 Oct 29;286(5441):958-61.
4. Wang C, Sanders CM, Yang Q, Schroeder HW, Wang E, Babrzadeh F, Gharizadeh B, Myers RM, Hudson JR, Davis RW, Han J. High throughput sequencing reveals a complex pattern of dynamic interrelationships among human T cell subsets. Proc Natl Acad Sci U S A. 2010 Jan 26;107(4):1518-23.
5. Robins HS, Srivastava SK, Campregher PV, Turtle CJ, Andriesen J, Riddell SR, Carlson CS, Warren EH. Overlap and effective size of the human CD8+ T-cell receptor repertoire. Sci Transl Med. 2010 Sep 1;2(47):47ra64.
6. Logan AC, Gao H, Wang C, Sahaf B, Jones CD, Marshall EL, Buto I, Armstrong R, Fire AZ, Weinberg KI, Mindrinos M, Zehnder JL, Boyd SD, Xiao W, Davis RW, Miklos DB. High-throughput VDJ sequencing for quantification of minimal residual disease in chronic lymphocytic leukemia and immune reconstitution assessment. Proc Natl Acad Sci U S A. 2011 Dec 27;108(52):21194-9.

7. Wu D, Sherwood A, Fromm JR, Winter SS, Dunsmore KP, Loh ML, Greisman HA, Sabath DE, Wood BL, Robins H. High-throughput sequencing detects minimal residual disease in acute T lymphoblastic leukemia. Sci Transl Med. 2012 May 16;4(134):134ra63.
8. Grupp SA, Kalos M, Barrett D, Aplene R, Porter DL, Rheingold SR, Teachey DT, Chew A, Hauck B, Wright JF, Milone MC, Levine BL, June CH. Chimeric Antigen Receptor- Modified T Cells for Acute Lymphoid Leukemia. N Engl J Med. 2013 Apr 18;368(16):1509-18.
9. Emerson RO, Sherwood AM, Rieder MJ, Guenthoer J, Williamson DW, Carlson CS, Drescher CW, Tewari M, Bielas JH, Robins HS. High-throughput sequencing of T-cell receptors reveals a homogeneous repertoire of tumour-infiltrating lymphocytes in ovarian cancer. J Pathol. 2013 Dec;231(4):433-40.
10. Zhu J, Peng T, Johnston C, Phasouk K, Kask AS, Klock A, Jin L, Diem K, Koelle DM, Wald A, Robins H, Corey L. Immune surveillance by CD8αα+ skin-resident T cells in human herpes virus infection. Nature. 2013 May 23;497(7450):494-7.
11. Luo W, Su J, Zhang X-B, Yang Z, Zhou M-Q, Jiang Z-M, Hao P-P, Liu S-D, Wen Q, Jin Q, Ma L. Limited T Cell Receptor Repertoire Diversity in Tuberculosis Patients Correlates with Clinical Severity. PLOS ONE. 2012 Oct 26;7(10):e48117.
12. Seay HR, Yusko E, Rothweiler SJ, Zhang L, Posgai AL, Campbell-Thompson M, Vignali M, Emerson RO, Kaddis JS, Ko D, Nakayama M, Smith MJ, Cambier JC, Pugliese A, Atkinson MA, Robins HS, Brusko TM. Tissue distribution and clonal diversity of the T and B cell repertoire in type 1 diabetes. JCI Insight [Internet]. [cited 2017 Oct 23];1(20). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5135280/
13. Rossetti M, Spreafico R, Consolaro A, Leong JY, Chua C, Massa M, Saidin S, Magni- Manzoni S, Arkachaisri T, Wallace CA, Gattorno M, Martini A, Lovell DJ, Albani S. TCR repertoire sequencing identifies synovial Treg cell clonotypes in the bloodstream during active inflammation in human arthritis. Ann Rheum Dis. 2017 Feb 1;76(2):435-41.
14. van Heijst JWJ, Ceberio I, Lipuma LB, Samilo DW, Wasilewski GD, Gonzales AMR, Nieves JL, van den Brink MRM, Perales MA, Pamer EG. Quantitative assessment of T cell repertoire recovery after hematopoietic stem cell transplantation. Nat Med. 2013 Mar;19(3):372-7.
15. Jones JL, Thompson SAJ, Loh P, Davies JL, Tuohy OC, Curry AJ, Azzopardi L, Hill-Cawthorne G, Fahey MT, Compston A, Coles AJ. Human autoimmunity after lymphocyte depletion is caused by homeostatic T-cell proliferation. Proc Natl Acad Sci. 2013 Dec 10;110(50):20200-5.
16. Neller MA, Burrows JM, Rist MJ, Miles JJ, Burrows SR. High Frequency of Herpesvirus- Specific Clonotypes in the Human T Cell Repertoire Can Remain Stable over Decades with Minimal Turnover. J Virol. 2013 Jan;87(1):697-700.

17. Yoshida K, Cologne JB, Cordova K, Misumi M, Yamaoka M, Kyoizumi S, Hayashi T, Robins H, Kusunoki Y. Aging-related changes in human T-cell repertoire over 20 years delineated by deep sequencing of peripheral T-cell receptors. Exp Gerontol. 2017 Oct 1;96(Supplement C):29-37.
18. Uematsu Y, Ryser S, Dembid Z, Borgulya P, Krimpenfort P, Berns A, von Boehmer H, Steinmetz M. In transgenic mice the introduced functional T cell receptor beta gene prevents expression of endogenous beta genes. Cell. 1988 Mar 25;52(6):831-41.
19. Woodsworth DJ, Castellarin M, Holt RA. Sequence analysis of T-cell repertoires in health and disease. Genome Med. 2013 Oct 30;5(10):98.
20. Glanville J, Huang H, Nau A, Hatton 0, Wagar LE, Rubelt F, Ji X, Han A, Krams SM, Pettus C, Haas N, Arlehamn CSL, Sette A, Boyd SD, Scriba TJ, Martinez OM, Davis MM. Identifying specificity groups in the T cell receptor repertoire. Nature. 2017 Jul 6;547(7661):94-8.
21. Emerson RO, DeWitt WS, Vignali M, Gravley J, Hu JK, Osborne EJ, Desmarais C, Klinger M, Carlson CS, Hansen JA, Rieder M, Robins HS. Immunosequencing identifies signatures of cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire. Nat Genet. 2017 Apr 3;49(5):ng.3822.
22. Wu D, Emerson RO, Sherwood A, Loh ML, Angiolillo A, Howie B, Vogt J, Rieder M, Kirsch I, Carlson C, Williamson D, Wood BL, Robins H. Detection of minimal residual disease in B lymphoblastic leukemia by high-throughput sequencing of IGH. Clin Cancer Res Off J Am Assoc Cancer Res. 2014 Sep 1;20(17):4540-8.
23. Sathaliyawala T, Kubota M, Yudanin N, Turner D, Camp P, Thome JJC, Bickham KL, Lerner H, Goldstein M, Sykes M, Kato T, Farber DL. Distribution and compartmentalization of human circulating and tissue-resident memory T cell subsets. Immunity. 2013 Jan 24;38(1):187-97.
24. Cossarizza A, Ortolani C, Paganelli R, Barbieri D, Monti D, Sansoni P, Fagiolo U, Castellani G, Bersani F, Londei M, Franceschi C. CD45 isoforms expression on CD4+ and CD8+ T cells throughout life, from newborns to centenarians: implications for T cell memory. Mech Ageing Dev. 1996 Mar 29;86(3):173-95.
25. Madi A, Poran A, Shifrut E, Reich-Zeliger S, Greenstein E, Zaretsky I, Arnon T, Laethem FV, Singer A, Lu J, Sun PD, Cohen IR, Friedman N. T cell receptor repertoires of mice and humans are clustered in similarity networks around conserved public CDR3 sequences. eLife. 2017 Jul 21;6.
26. Miho E, Greiff V, Roskar R, Reddy ST. The fundamental principles of antibody repertoire architecture revealed by large-scale network analysis. bioRxiv. 2017 Apr 5;124578.
27. Greiff V, Menzel U, Miho E, Weber C, Riedel R, Cook S, Valai A, Lopes T, Radbruch A, Winkler TH, Reddy ST. Systems Analysis Reveals High Genetic and

Antigen-Driven Predetermination of Antibody Repertoires throughout B Cell Development. Cell Rep. 2017 May 16;19(7):1467-78.
28. Madi A, Shifrut E, Reich-Zeliger S, Gal H, Best K, Ndifon W, Chain B, Cohen IR, Friedman N. T-cell receptor repertoires share a restricted set of public and abundant CDR3 sequences that are associated with self-related immunity. Genome Res. 2014 Jul 14;gr.170753.113.
29. Elhanati Y, Murugan A, Callan CG, Mora T, Walczak AM. Quantifying selection in immune receptor repertoires. Proc Natl Acad Sci. 2014 Jul 8;111(27):9875-80.
30. Murugan A, Mora T, Walczak AM, Callan CG. Statistical inference ofthe generation probability of T-cell receptors from sequence repertoires. Proc Natl Acad Sci U S A. 2012 Oct 2;109(40):16161-6.
31. Venturi V, Quigley MF, Greenaway HY, Ng PC, Ende ZS, McIntosh T, Asher TE, Almeida JR, Levy S, Price DA, Davenport MP, Douek DC. A mechanism for TCR sharing between T cell subsets and individuals revealed by pyrosequencing. J Immunol Baltim Md 1950. 2011 Apr 1;186(7):4285-94.
32. Klarenbeek PL, Remmerswaal EBM, Berge IJM ten, Doorenspleet ME, Schaik BDC van, Esveldt REE, Koch SD, Brinke A ten, Kampen AHC van, Bemelman FJ, Tak PP, Baas F, Vries N de, Lier RAW van. Deep Sequencing of Antiviral T-Cell Responses to HCMV and EBV in Humans Reveals a Stable Repertoire That Is Maintained for Many Years. PLOS Pathog. 2012 Sep 27;8(9):e1002889.
33. Marcou Q, Mora T, Walczak AM. IGoR: a tool for high-throughput immune repertoire analysis. ArXiv170508246 Q-Bio [Internet]. 2017 May 23 [cited 2018 Feb 7]; Available from: http://arxiv.org/abs/1705.08246
34. Qi Q, Liu Y, Cheng Y, Glanville J, Zhang D, Lee J-Y, Olshen RA, Weyand CM, Boyd SD, Goronzy JJ. Diversity and clonal selection in the human T-cell repertoire. Proc Natl Acad Sci U S A. 2014 Sep 9;111(36):13139-44.
35. Greiff V, Bhat P, Cook SC, Menzel U, Kang W, Reddy ST. A bioinformatic framework for immune repertoire diversity profiling enables detection of immunological status. Genome Med. 2015 May 28;7(1):49.
36. Dash P, Fiore-Gartland AJ, Hertz T, Wang GC, Sharma S, Souquette A, Crawford JC, Clemens EB, Nguyen THO, Kedzierska K, La Gruta NL, Bradley P, Thomas PG. Quantifiable predictive features define epitope-specific T cell receptor repertoires. Nature. 2017 Jul 6;547(7661):89-93.
37. Matos TR, O'Malley JT, Lowry EL, Hamm D, Kirsch IR, Robins HS, Kupper TS, Krueger JG, Clark RA. Clinically resolved psoriatic lesions contain psoriasis-specific IL-17- producing αβ T cell clones. J Clin Invest. 2017 Nov 1;127(11):4031-41.
38. Leeuwen EMM van, Remmerswaal EBM, Heemskerk MHM, Berge IJM ten, Lier RAW van. Strong selection of virus-specific cytotoxic CD4+ T-cell clones during primary human cytomegalovirus infection. Blood. 2006 Nov 1;108(9):3121-7.

39. Snyder A, Nathanson T, Funt SA, Ahuja A, Novik JB, Hellmann MD, Chang E, Aksoy BA, Al-Ahmadie H, Yusko E, Vignali M, Benzeno S, Boyd M, Moran M,Lyer G, Robins HS, Mardis ER, Merghoub T, Hammerbacher J, Rosenberg JE, Bajorin DF. Contribution of systemic and somatic factors to clinical response and resistance to PD-L1 blockade in urothelial cancer: An exploratory multi-omic analysis. PLOS Med. 2017 May 26;14(5):e1002309.
40. Page DB, Yuan J, Redmond D, Wen YH, Durack JC, Emerson R, Solomon S, Dong Z, Wong P, Comstock C, Diab A, Sung J, Maybody M, Morris E, Brogi E, Morrow M, Sacchini V, Elemento O, Robins H, Patil S, Allison JP, Wolchok JD, Hudis C, Norton L, McArthur HL. Deep Sequencing of T-cell Receptor DNA as a Biomarker of Clonally Expanded TILs in Breast Cancer after Immunotherapy. Cancer Immunol Res. 2016 Oct;4(10):835-44.
41. Meier J, Roberts C, Avent K, Hazlett A, Berrie J, Payne K, Hamm D, Desmarais C, Sanders C, Hogan KT, Archer KJ, Manjili MH, Toor AA. Fractal organization of the human T cell repertoire in health and after stem cell transplantation. Biol Blood Marrow Transplant J Am Soc Blood Marrow Transplant. 2013 Mar;19(3):366-77.
42. Warren RL, Freeman JD, Zeng T, Choe G, Munro S, Moore R, Webb JR, Holt RA. Exhaustive T-cell repertoire sequencing of human peripheral blood samples reveals signatures of antigen selection and a directly measured repertoire size of at least 1 million clonotypes. Genome Res. 2011 May;21(5):790-7.
43. Klein SL, Flanagan KL. Sex differences in immune responses. Nat Rev Immunol. 2016 Oct;16(10):626-38.
44. Weiskopf D, Weinberger B, Grubeck-Loebenstein B. The aging ofthe immune system. Transpl Int. 2009 Nov 1;22(11):1041-50.
45. Lefranc M-P, Giudicelli V, Ginestoux C, Jabado-Michaloud J, Folch G, Bellahcene F, Wu Y, Gemrot E, Brochet X, Lane J, Regnier L, Ehrenmann F, Lefranc G, Duroux P. IMGT, the international ImMunoGeneTics information system. Nucleic Acids Res. 2009 Jan;37(Database issue):D1006-1012.
46. Carlson CS, Emerson RO, Sherwood AM, Desmarais C, Chung M-W, Parsons JM, Steen MS, LaMadrid-Herrmannsfeldt MA, Williamson DW, Livingston RJ, Wu D, Wood BL, Rieder MJ, Robins H. Using synthetic templates to design an unbiased multiplex PCR assay. Nat Commun. 2013;4:2680.

Figures

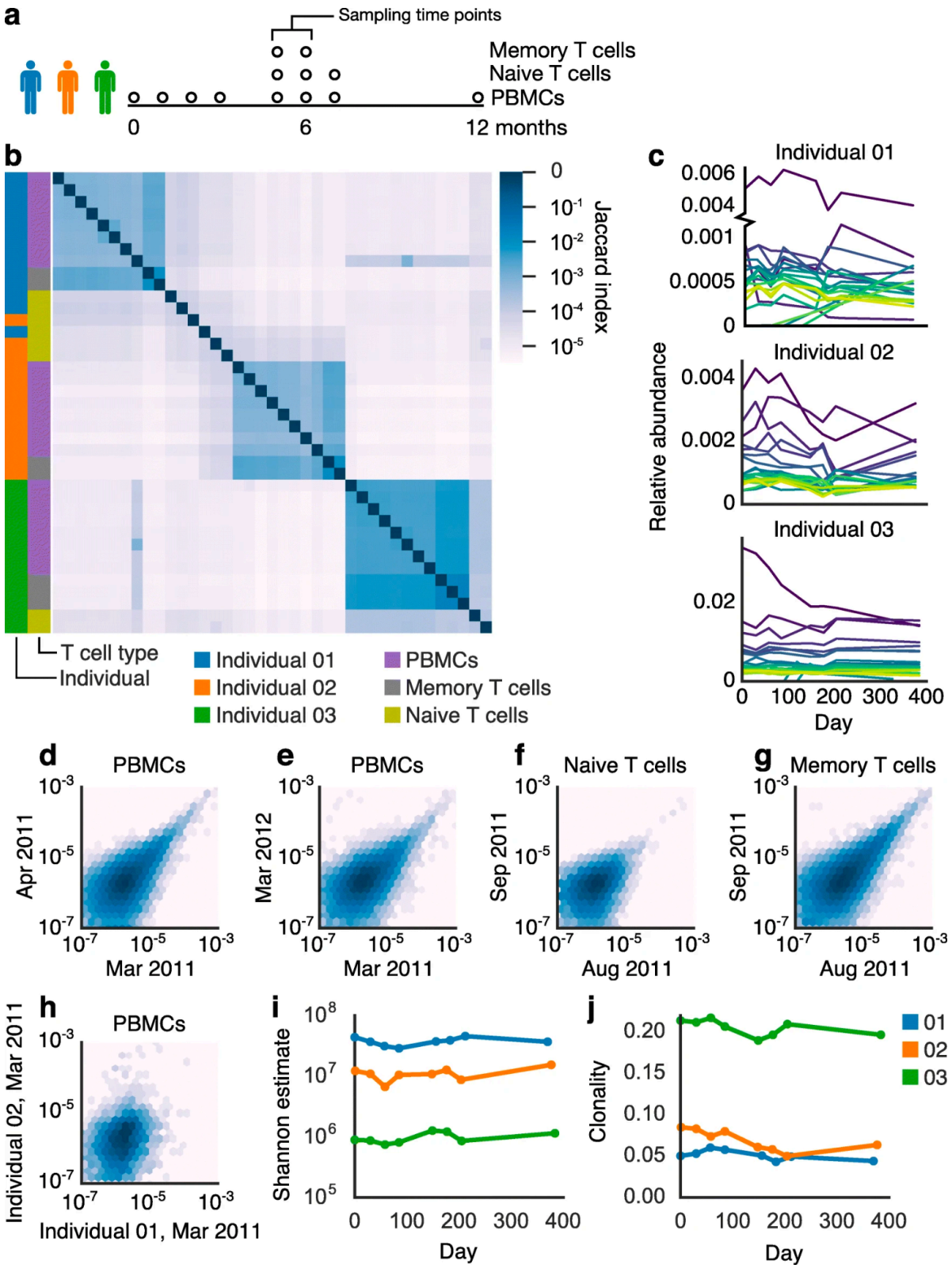

a
Sampling time points
Memory T cells
Naive T cells
PBMCs
0
6
12 months
b
Jaccard index
T cell type
Individual
Individual 01
Individual 02
Individual 03
PBMCs
Memory T cells
Naive T cells
c
Individual 01
Individual 02
Individual 03
Relative abundance
Day
d
PBMCs
Apr 2011
Mar 2011
e
PBMCs
Mar 2012
Mar 2011
f
Naive T cells
Sep 2011
Aug 2011
g
Memory T cells
Sep 2011
Aug 2011
h
PBMCs
Individual 02, Mar 2011
Individual 01, Mar 2011
i
Shannon estimate
Day
j
Clonality
Day
01
02
03

Figure 1. The TCRs repertoire displayed stability and individual-specific characteristics across time. (**a**) Experimental design of T cell sampling. (**b**) A heatmap of Jaccard indexes shows clear clustering of samples by individual. Samples of naive T cells clustered less by individual than did PBMC or memory T cell samples. Relative abundances of the 20 most abundant TCRβs (**c**) appeared stable through time. TCRβ abundances in PBMCs correlated within an individual across time points, including across a month (**d,** shared TCRβs = 33601, Spearman *rho* = 0.55718, $p < 10^{-6}$**),** and a year (e, shared TCRβ*s* = *25933,* Spearman *rho* = 0.53810, $p < 10^{-6}$), as well as across a month in naive (**f,** shared TCRβs = 15873, Spearman *rho* = 0.37892, $p < 10^{-6}$) and memory T cells (**g,** shared TCRβs = 47866, Spearman *rho* **=** 0.64934, $p < 10^{-6}$). TCRβs correlated much less across individuals (**h**, shared TCRβs = 5014, Spearman *rho* = 0.28554, $p < 10^{-6}$). Shannon alpha diversity estimate (**i**) and clonality (defined as 1 **-** Pielou's evenness, **j**) of the TCRβ repertoire were consistent over time.

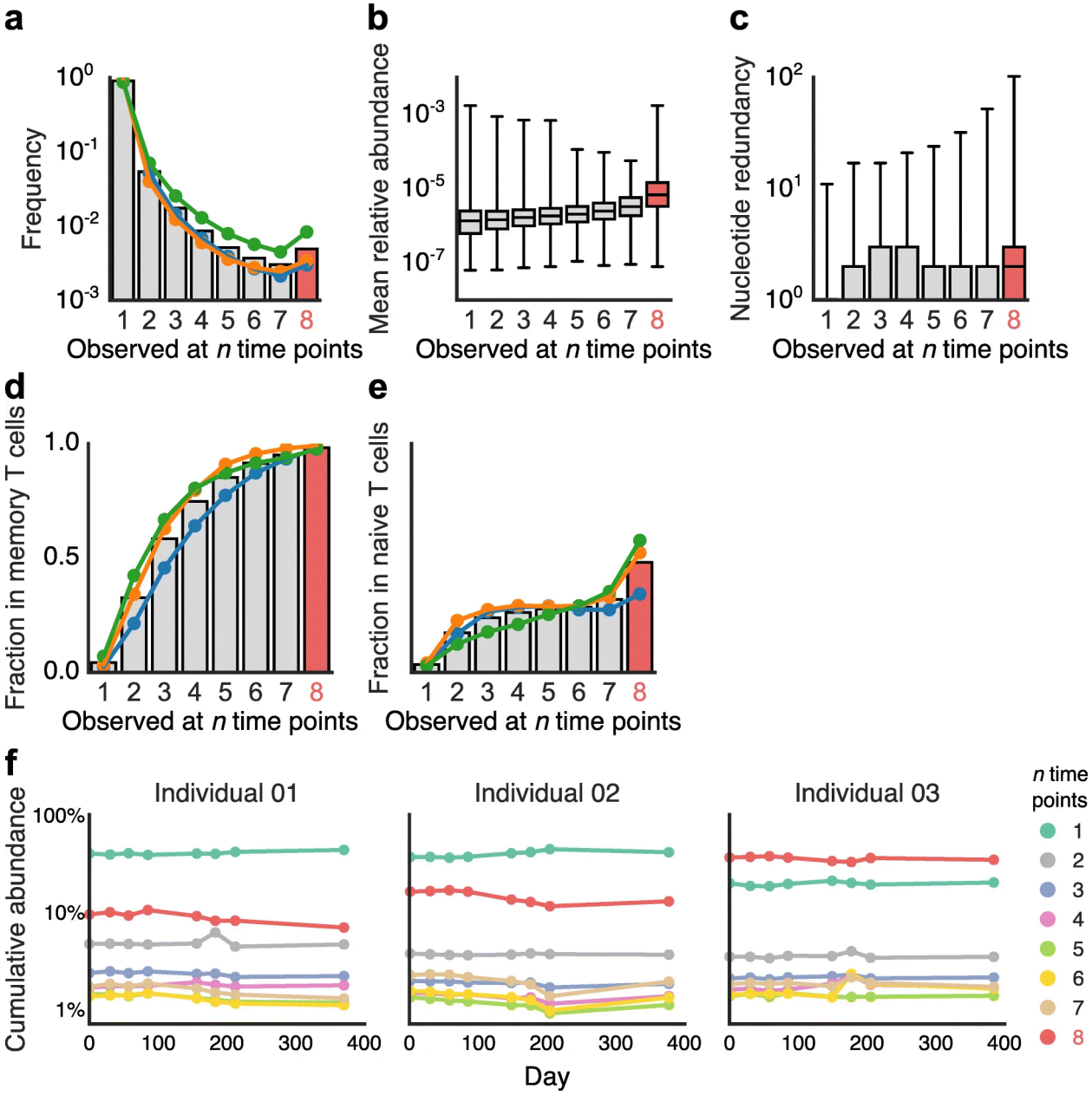


Figure 2. A subset of the TCRβ repertoire occurred across all time points-the persistent TCRβ repertoire. (**a**) The number of TCRβs observed at n time points. Persistent TCRβs tended to have (**b**) greater abundance (Mann-Whitney *U* test, statistic=26297052589.5, $p < 10^{-308}$) and (**c**) nucleotide sequence redundancy (Mann-Whitney *U* test, statistic=25851211348.0, $p < 10^{-308}$) than other receptors. Mann-Whitney *U* tests between groups are in Tables S6, S7. Persistent TCRβs had higher proportions of TCRβs in common with memory (**d**) and with naive (**e**) T cell populations and constituted a stable and significant fraction of overall TCRβ abundance across time (**f**).

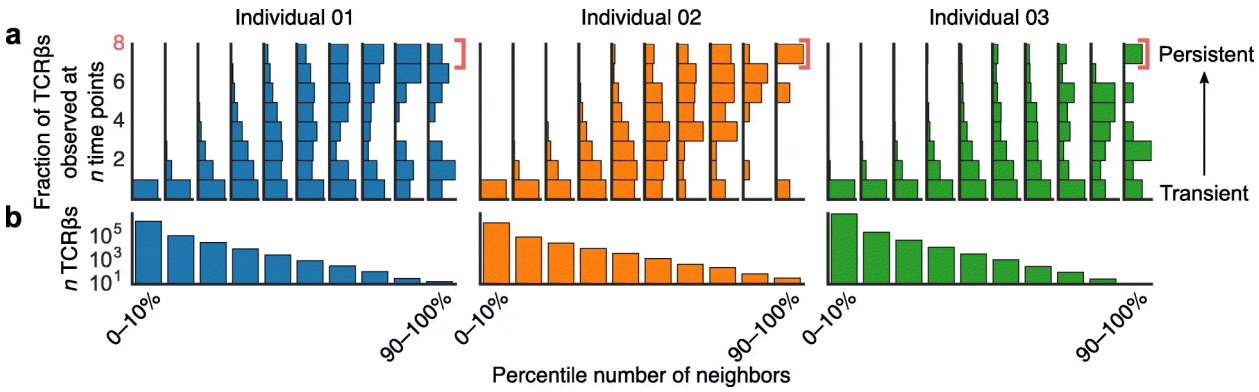


Figure 3. Persistent TCRβs were more functionally redundant. We created a network graph of TCRβs from each individual, drawing edges between TCRβs on the basis of sequence similarity (Levenshtein distances), which reflects antigen specificity. We then grouped TCRβs into decile bins based on the number of neighbors (similar TCRβs) of each TCRβ. In other words, TCRβs in the 0-10% bin had 0% to 10% of the maximum number of neighbors observed for any TCRβ – the fewest neighbors – while those in the 90-100% bin had near the maximum number of neighbors observed. For each decile bin, we then counted how many samples each TCRβ occurred in from our time series data. (a) Vertical histograms of these distributions indicate that TCRβs with few neighbors -and thus few similar observed TCRβs-tended to occur at only a single time point, while TCRβs with more neighbors-and thus higher numbers of similar TCRβs observed-tended to have a higher proportion of persistent TCRβs. (**b**) The number of TCRβs in each neighbor bin (Figure S13a).

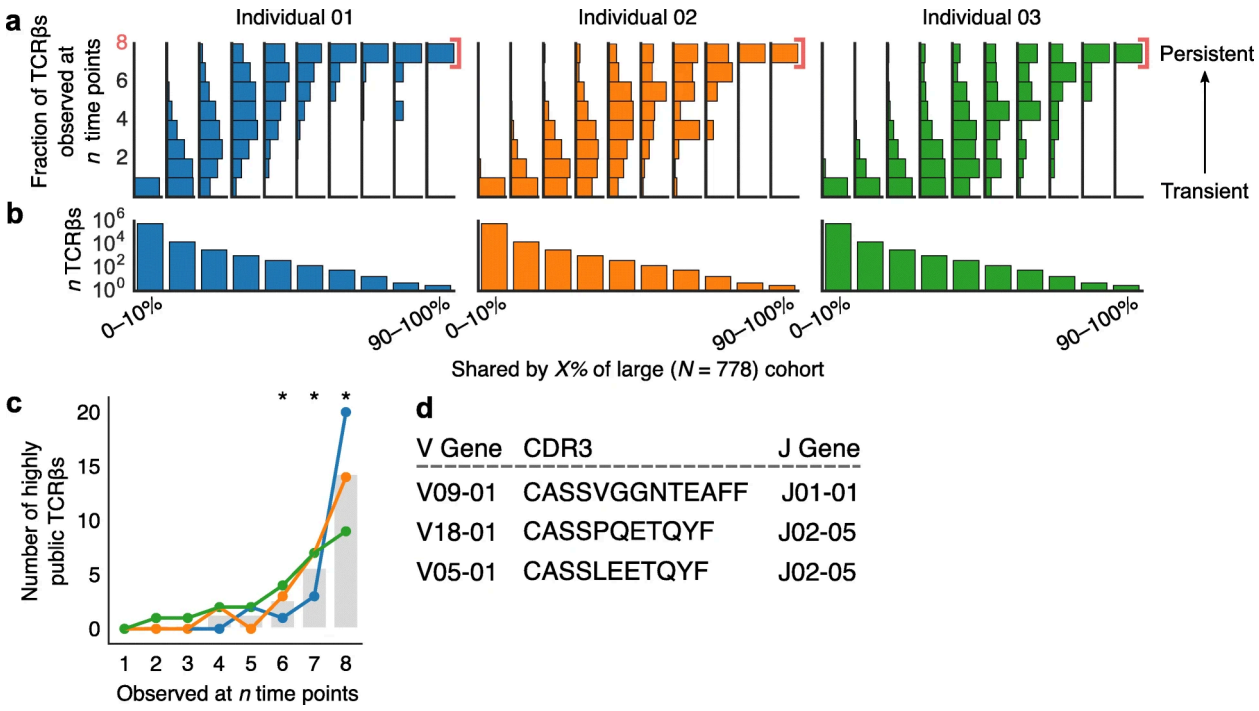


Figure 4. Persistent TCRβs were enriched in highly public TCRβs. We identified public TCRs occurring in 0-10%, 0-20%,. .. 90-100% of individuals in an independent, large cohort of similarly profiled subjects (*N* = 778). For each of these decile bins, we examined TCRβs shared across each of our three individuals' time series data and tallied the number of time points at which we observed each TCRβ. (**a**) Vertical histograms of these distributions indicate that more- private TCRβs-TCRβs shared by few people-occurred most often at only a single time point, while more-public TCRβs tended to persist across time. (**b**) The number of TCRβs evaluated in each decile bin. The vast majority of receptors were not shared or were shared across few individuals (also see Figure S13b). (**c**) In all three individuals in this study, persistent TCRβs included greater numbers of highly public TCRβs-defined here as receptors shared by over 70% of subjects from the large cohort-than receptors that only occurred once (independent t-test, statistic=-4.508, *p*=0.01). Asterisks indicate $p < 0.05$. (**d**) The three most public TCRβs (in over 90% of 778 individuals) were also persistent in all three individuals.

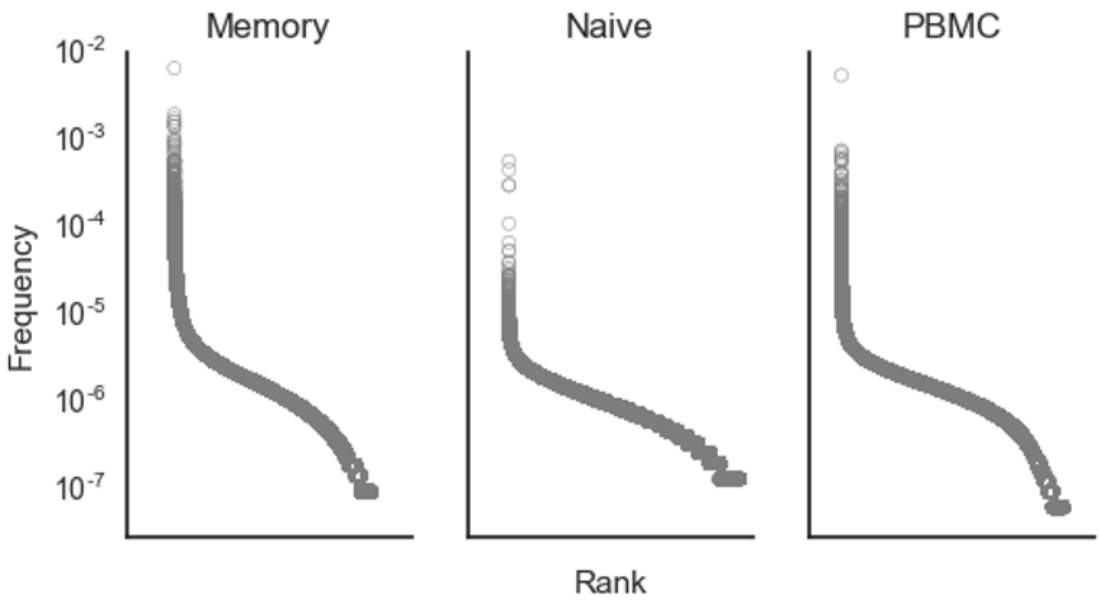


Figure S1. Representative frequency rank plots for memory T cells, naive T cells, and all T cells from PBMCs from Individual 01. As expected, naive T cells had fewer abundant clones than PBMC or memory T cells. In all cases, the majority of TCRβs had abundances around 10^{-6}.

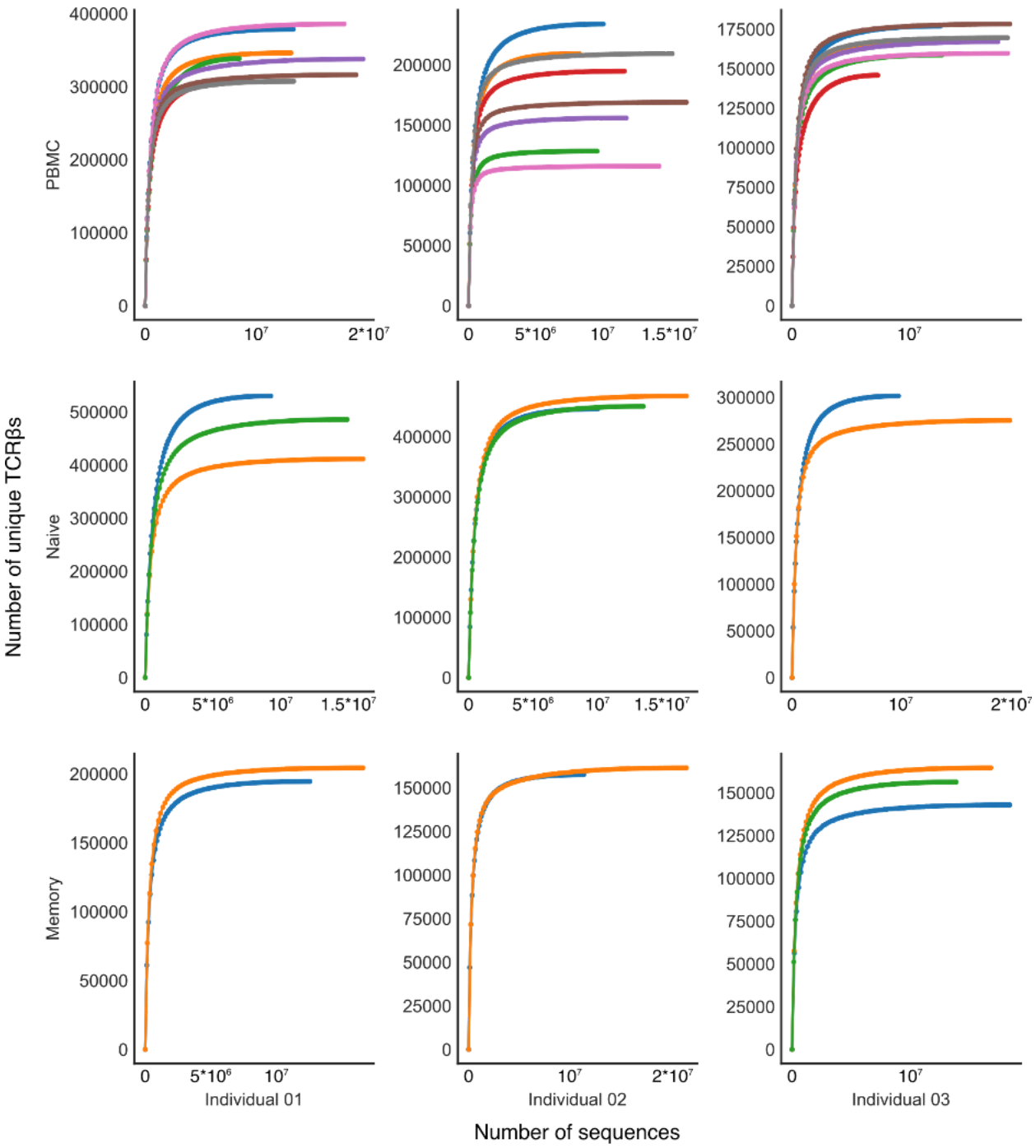


Figure S2. Rarefaction curves for each subject indicate that sample libraries were sequenced well past saturation.

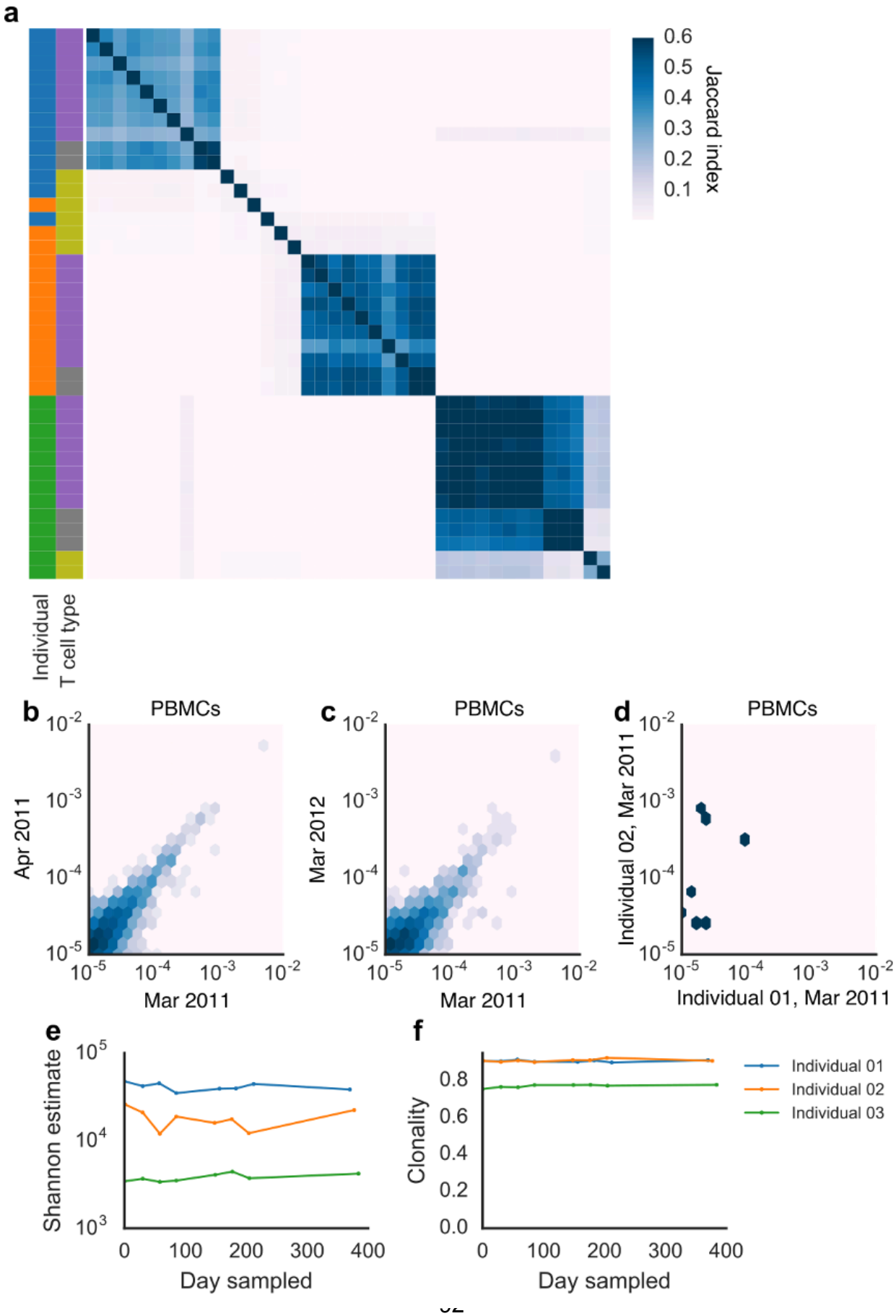

a
0.6
0.5
0.4
0.3
0.2
0.1
Jaccard index
Individual
T cell type
b
PBMCs
Apr 2011
Mar 2011
c
PBMCs
Mar 2012
Mar 2011
d
PBMCs
Individual 02, Mar 2011
Individual 01, Mar 2011
e
Shannon estimate
Day sampled
f
Clonality
Day sampled
Individual 01
Individual 02
Individual 03

Figure S3. Analyses examining only high-abundance TCRβs agree with results from full- repertoire analysis, suggesting that undersampling likely did not confound our results. (**a**) A heatmap of Jaccard indexes shows similar clustering of PBMC and memory T cell samples by individual and less clustering of naive T cell samples. Abundances of high-abundance TCRβs in PBMC samples correlated within an individual (individual 01) across time points, including across a month (**b,** shared TCRβ = 2057, Spearman *rho* = 0.66902, $p < 10^{-6}$) and a year (**c**, shared TCRβs = 1390, Spearman *rho* = 0.59251, $p < 10^{-6}$). High-abundance TCRβs did not appear to correlate across individuals, largely because of lack of shared TCRβs (**d,** shared TCRβs = 7, Spearman *rho* **=** 0.14286, *p = 0.75995).* Shannon alpha diversity estimate (**e**) and clonality (defined as 1 - Pielou's evenness, **f**) of the TCRβ repertoire were consistent over time.

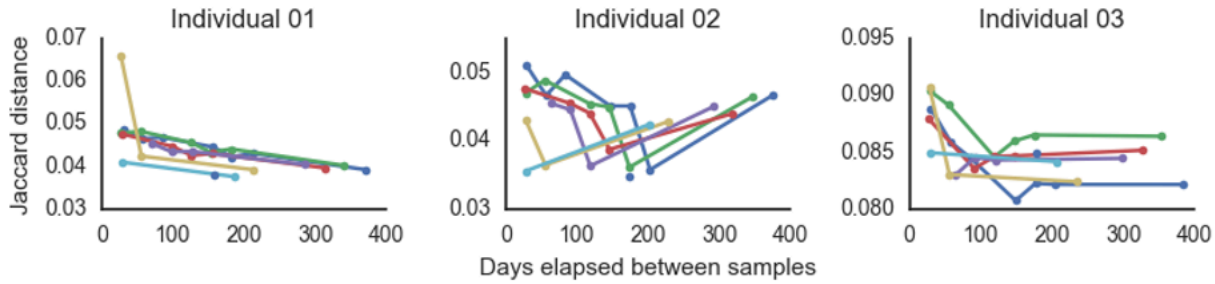


Figure S4. TCRβ repertoire overlap (Jaccard index) often decreases with increasing time between samples, except in Individual 02, where the final time point at one year past the first sample shared more TCRβs with the previous samples. Different colors depict changes over time relative to each sample. For example, the blue line depicts overlap between each sample after the first sample and the first sample, while the green line depicts overlap between each sample after the second sample and the second sample.

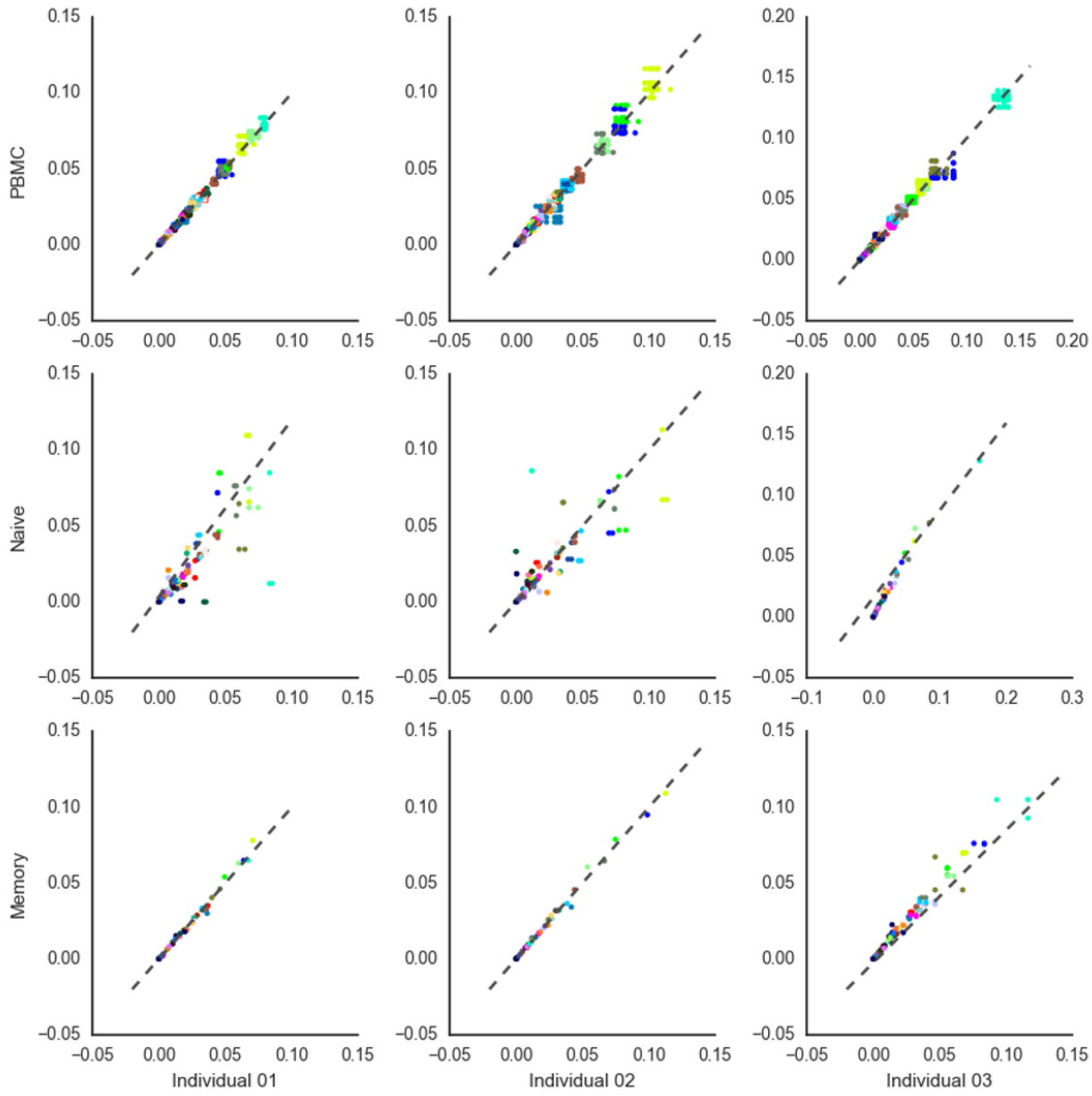


Figure S5. V gene usage across time and cell compartment in all three individuals. Each of nine plots represents each cell type in each individual. Within each plot, different colors represent different V genes, and each dot represents a comparison of the abundances of that V gene from one sample to another. Points that fall near a 1:1 ratio (indicated by the dotted line) are nearly identical in abundance between the two samples considered. These plots indicate that VJ gene usage was generally the same across time points, particularly in total and memory T cells. In naive cells, VJ gene usage varied more.

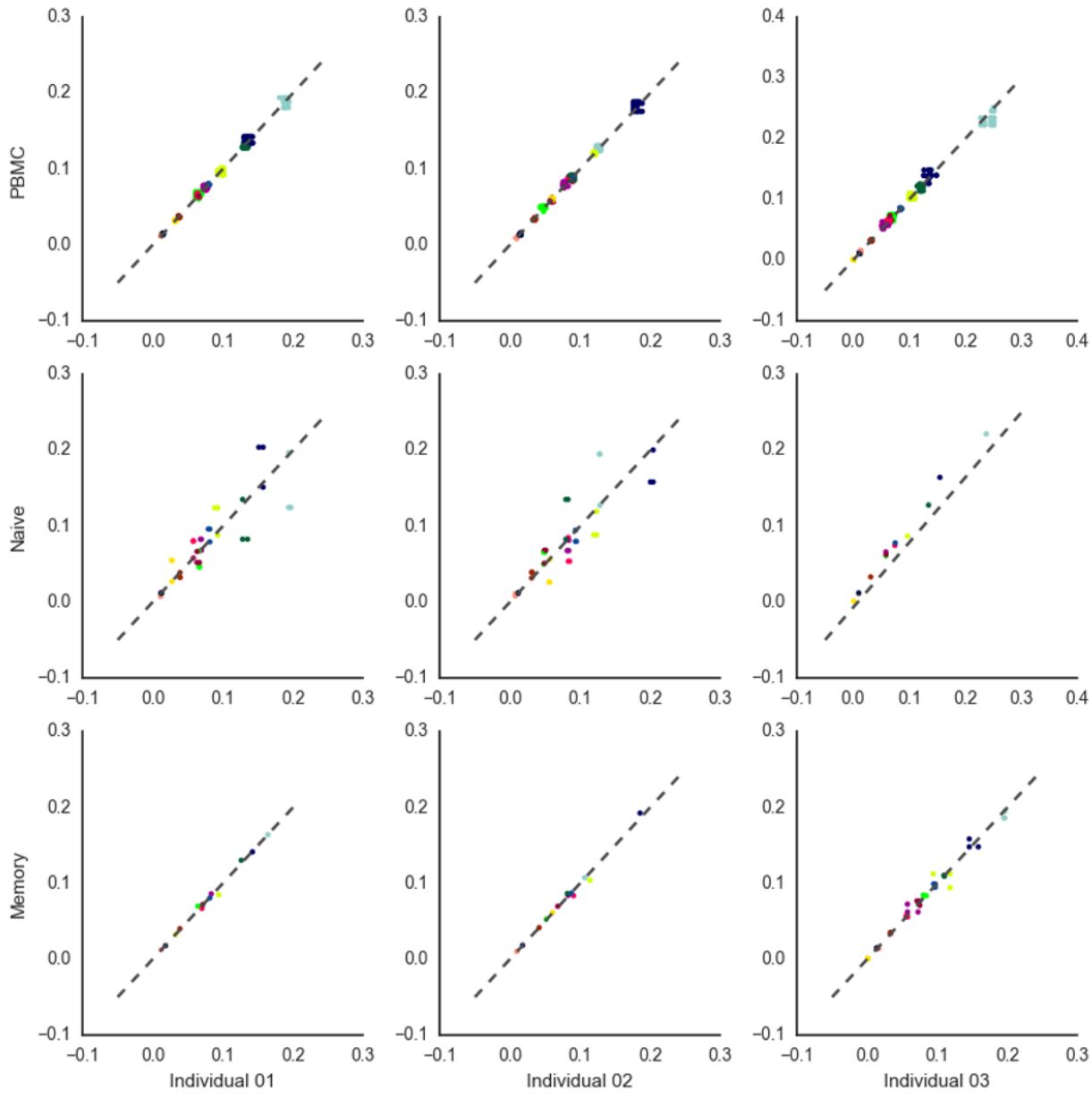


Figure S6. J gene usage across time and cell compartment in all three individuals. Plots are as in Figure S5, with similar findings.

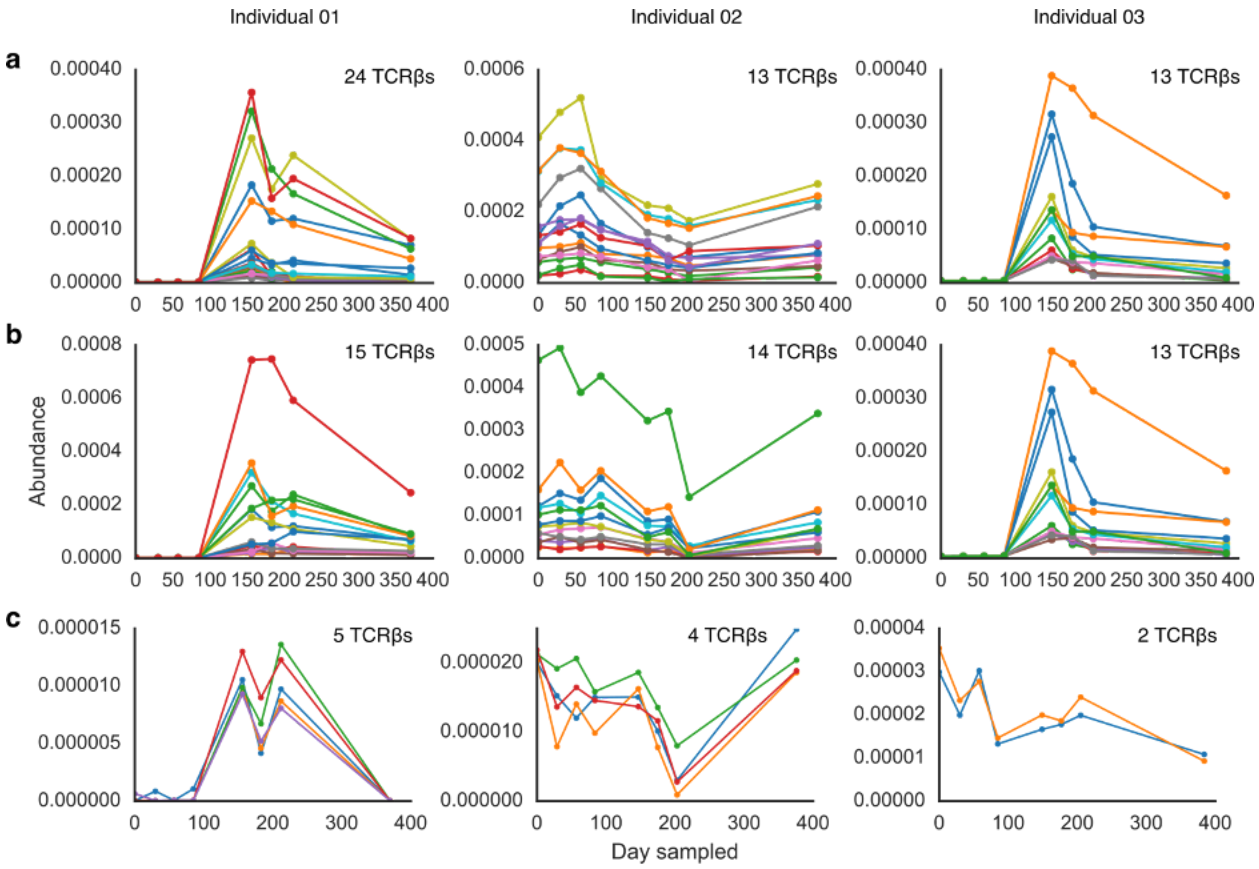


Figure S7. Cohorts of TCRβs exhibit correlated dynamics over time. We found large cohorts of correlating TCRβs by Spearman (**a**) and Pearson (**b**) correlation. Although these TCRβs spanned a range of abundances, we did not observe any clear signs of correlation caused by sequencing or library preparation errors (Table S2). We also found smaller cohorts (**c**) of TCRβs with nearly identical abundances whose dynamics also correlated through time. The number of TCRβs found in all cohorts was significant ($p < 0.001$) in a random permutation test (see Methods). These TCRβ cohorts might be an artifact of sampling noise, or they may represent receptors involved in the same immune response.

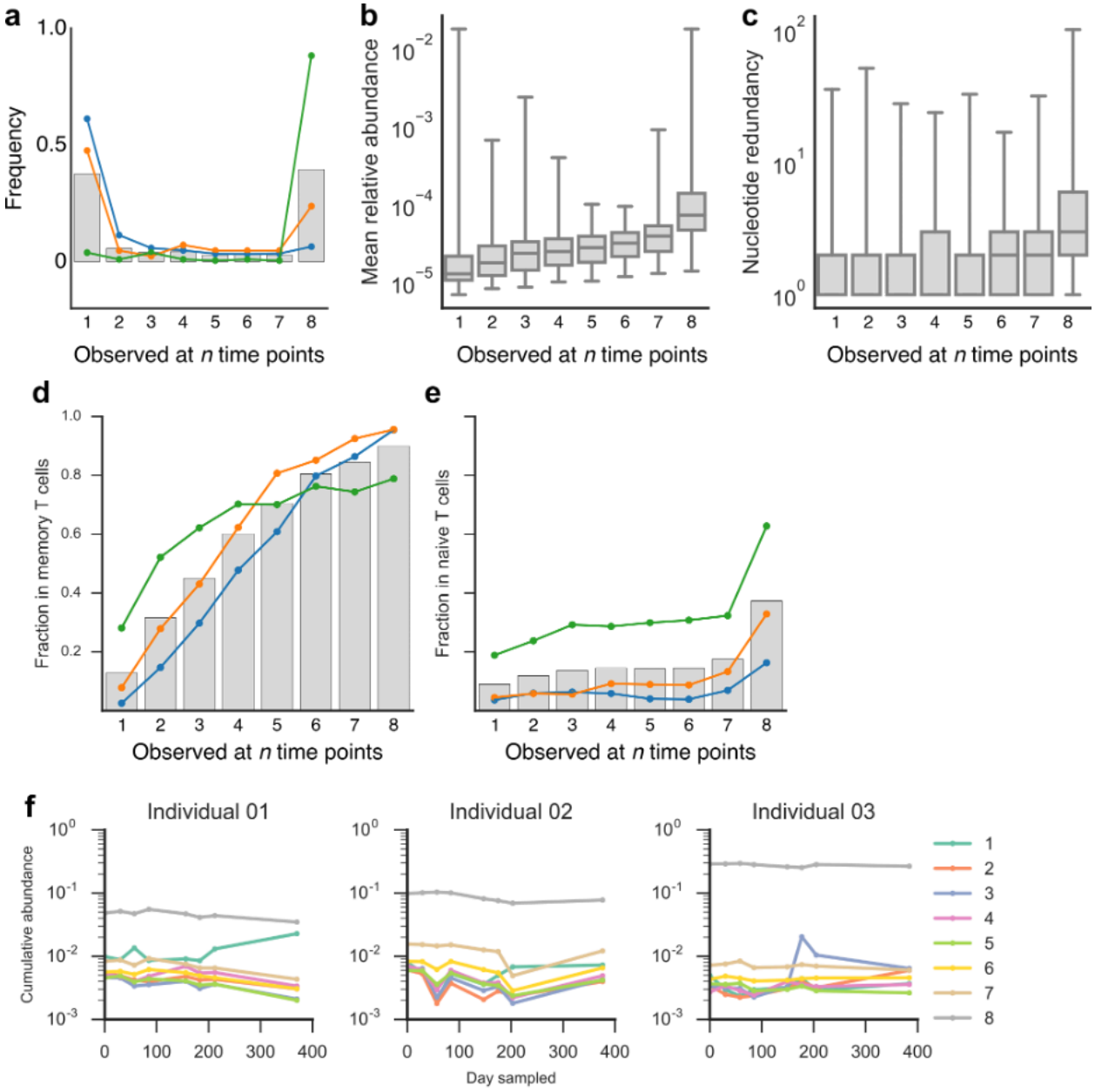


Figure S8. Persistent high-abundance TCRβs exhibit similar patterns as overall persistent TCRβs. (**a**) High-abundance TCRβs had a greater prevalence of persistent TCRβs, although the exact values varied across individuals. Persistent high-abundance TCRβs also showed greater mean abundance (**b**) and nucleotide redundancy (**c**). Persistent high-abundance TCRβs also had higher proportions of TCRβs in common with memory (**d**) and naive (**e**) T cell populations and constituted a stable and significant fraction of overall TCRβ abundance across time (**f**).

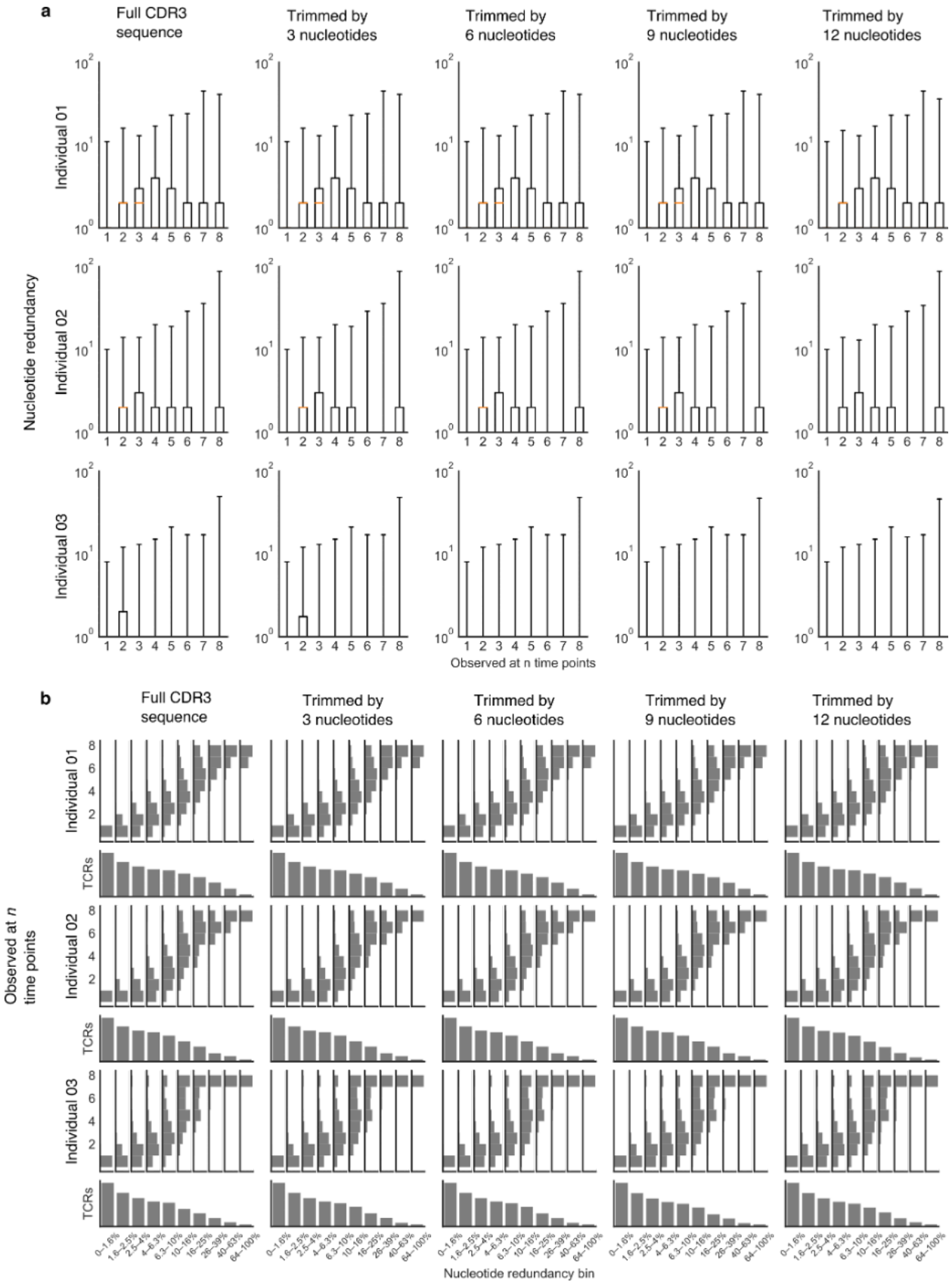

a
Full CDR3 sequence
Trimmed by 3 nucleotides
Trimmed by 6 nucleotides
Trimmed by 9 nucleotides
Trimmed by 12 nucleotides
Individual 01
Individual 02
Individual 03
Nucleotide redundancy
Observed at n time points
b
Observed at n time points
TCRs
0–1.6%
1.6–2.5%
2.5–4%
4–6.3%
6.3–10%
10–16%
16–25%
25–39%
40–63%
64–100%
Nucleotide redundancy bin

Figure S9. Nucleotide redundancy across individuals and with more stringent assignment of CDR3 sequence (figure supplements Figure 2c). (**a**) Each plot represents nucleotide redundancy for TCRβs that were observed in n samples. Rows represent plots for each individual. The left- most column of plots comprises data from full CDR3 nucleotide sequences as identified by IMGT (as in Figure 2c): we observed that the pattern of increasing nucleotide redundancy in persistent TCRβs was not consistent across individuals. Each of the following columns plot data from CDR3 nucleotide sequences that were progressively trimmed on each end by 3, 6, 9, and 12 nucleotides. We trimmed these sequences because CDR3 sequences identified by IMGT generally capture a number of amino acids – usually one to four at each end of the sequence – that are derived from V and J genes. Nucleotide mutations in these leading and trailing ends are thus less likely to be of biological origin and more likely to be from sequencing error, since we do not expect nucleotides from the V or J genes to be altered during TCR recombination (except for deletions). From these plots, we can observe that nucleotide redundancy is generally stable over different lengths of trimming, suggesting that our data are not skewed by these potential sequencing errors. (**b**) To further examine the relationship between persistence and nucleotide redundancy, we grouped TCRβs into 10 bins according to nucleotide redundancy. Because nucleotide redundancy is extremely skewed-the vast majority of TCRβs are encoded by a single clonotype-we created these bins on a logarithmic scale: the first bin includes TCRβs with nucleotide redundancy values up to 1.6% of the maximum value for each individual; the second between 1.6% and *2.5%* of the maximum value; and up to the 10th bin, which includes TCRβs with nucleotide redundancy values between 64% and 100% of the maximum value. For each of these TCRβ bins, we then plotted a histogram of the frequency of TCRβs that were observed at n time points. We observe a clear pattern across individuals and trimming lengths: TCRβs with greater nucleotide redundancy tend to occur at more time points, and the most redundant TCRβs are exclusively persistent receptors.

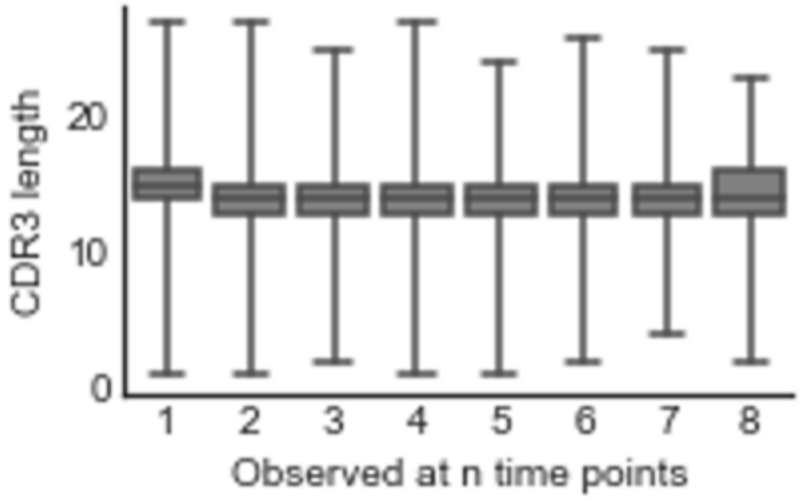


Figure S10. The persistent TCR repertoire exhibited little alteration of CDR3 lengths.

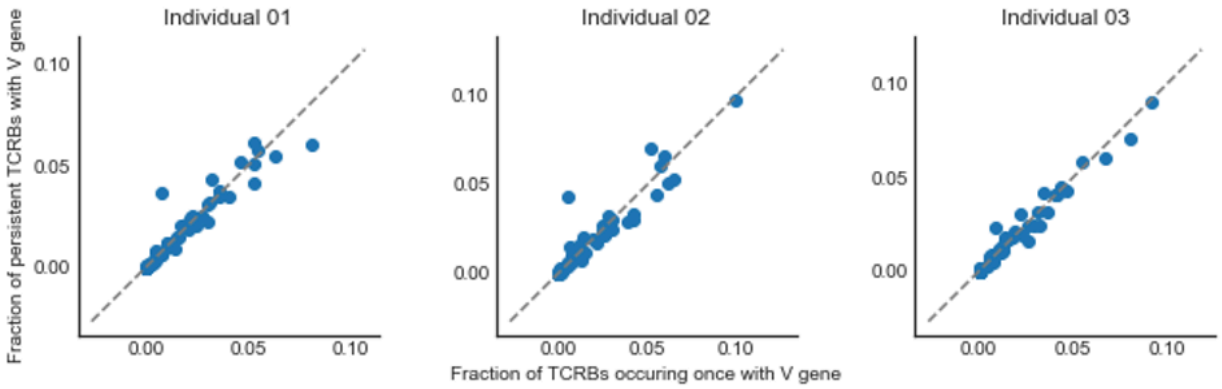


Figure S11. The persistent TCRβ repertoire does not exhibit altered V gene usage. These plots show V gene usage in TCRβs that occurred only once (x-axis) versus in persistent TCRβs (y axis). Each data point represents a single V gene. These values were closely correlated.

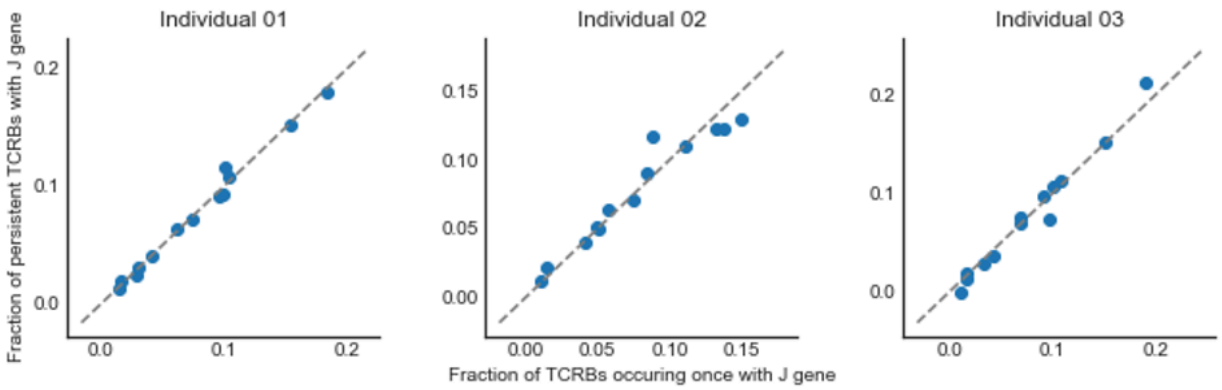


Figure S12. The persistent TCRβ repertoire does not exhibit altered J gene usage. Similar plots as in Figure S10 indicate that J gene usage is not greatly changed in persistent TCRβs.

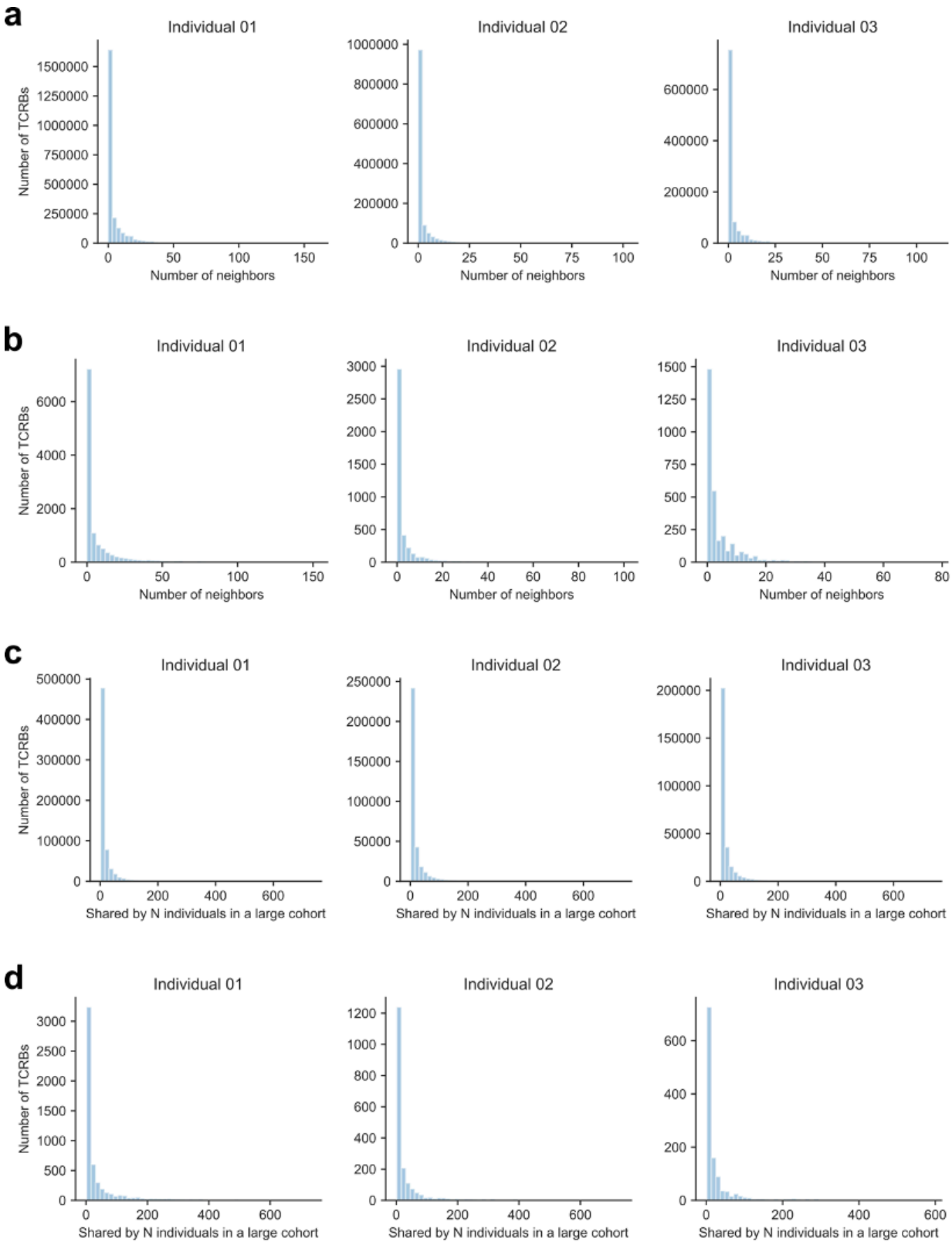

a
Individual 01
Individual 02
Individual 03
Number of TCRBs
Number of neighbors
b
Individual 01
Individual 02
Individual 03
Number of TCRBs
Number of neighbors
c
Individual 01
Individual 02
Individual 03
Number of TCRBs
Shared by N individuals in a large cohort
d
Individual 01
Individual 02
Individual 03
Number of TCRBs
Shared by N individuals in a large cohort

Figure S13. Distributions of the number of neighbors and degree of sharing across people for all TCRβs and high-abundance TCRβs. Each plot is a distribution of all TCRβs from PBMC samples for each individual (**a**, **c**) or only high-abundance TCRβs (**b, d**). Plots (**a**) and (**b**) show the number of neighbors in a network based on Levenshtein distance and plots (**c**) and (**d**) show the number of subjects sharing a given receptor in a large, independent, and similarly profiled cohort.

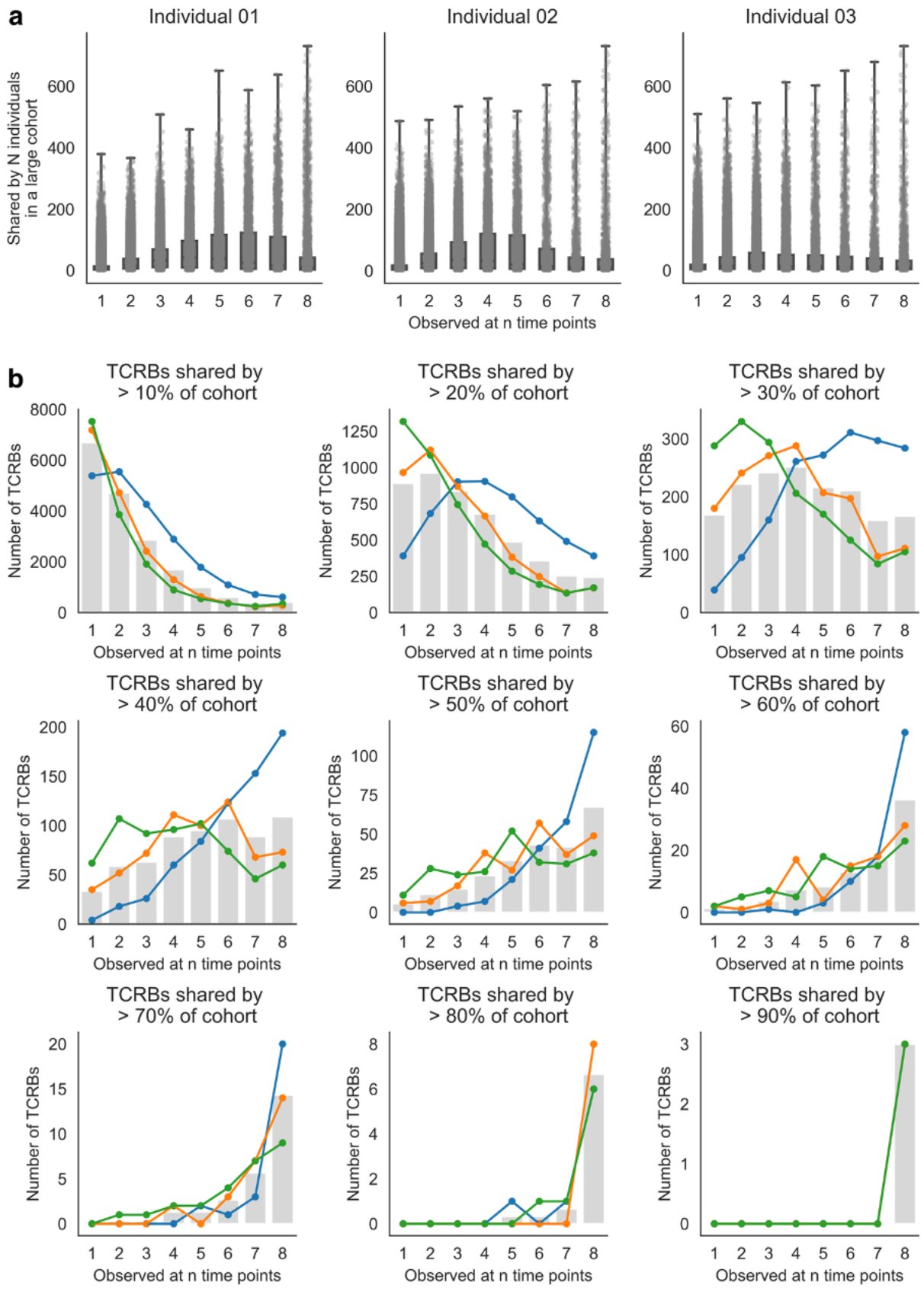
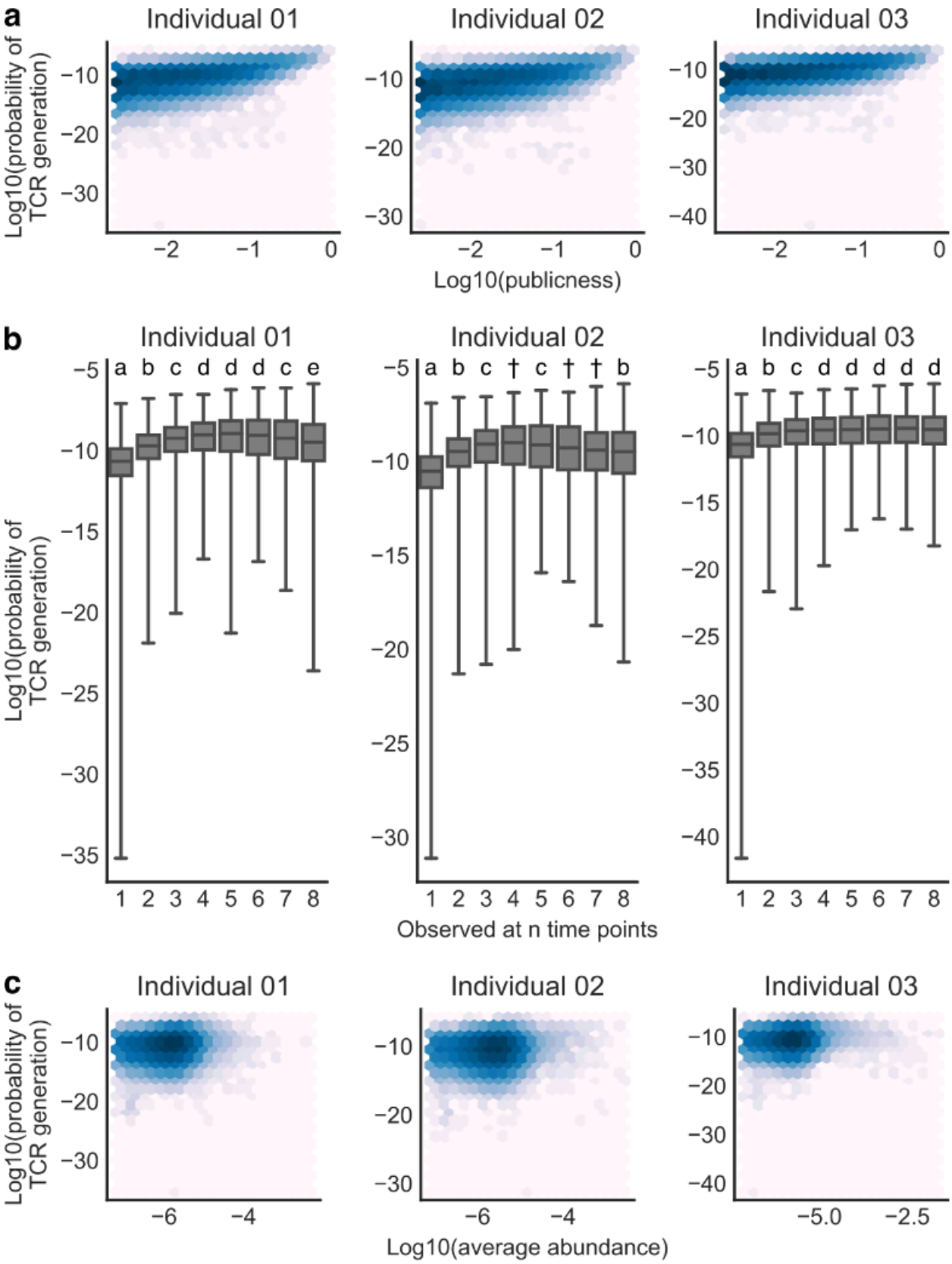

a
Individual 01
Individual 02
Individual 03
Shared by N individuals in a large cohort
Observed at n time points
b
TCRBs shared by > 10% of cohort
TCRBs shared by > 20% of cohort
TCRBs shared by > 30% of cohort
TCRBs shared by > 40% of cohort
TCRBs shared by > 50% of cohort
TCRBs shared by > 60% of cohort
TCRBs shared by > 70% of cohort
TCRBs shared by > 80% of cohort
TCRBs shared by > 90% of cohort
Number of TCRBs
Observed at n time points

Figure S14. Persistent TCRβs were rich in highly public TCRβs. (**a**) Over all TCRβs, receptors that occurred at an intermediate number of time points were on average most-shared across people, but these distributions are heavily skewed toward private receptors. (**b**) We focused on TCRβs that were shared by at least a certain percentage of individuals in the large (*N* = 778) cohort. We found that less public TCRβs were generally observed at few time points, while highly public TCRβs were predominantly observed at all time eight points. These results were even more striking given that we observed ~100-1000-fold more TCRβs occurring at a single time point than at all time points.


a
Individual 01
Individual 02
Individual 03
Log10(probability of TCR generation)
Log10(publicness)
b
Individual 01
Individual 02
Individual 03
Log10(probability of TCR generation)
Observed at n time points
c
Individual 01
Individual 02
Individual 03
Log10(probability of TCR generation)
Log10(average abundance)

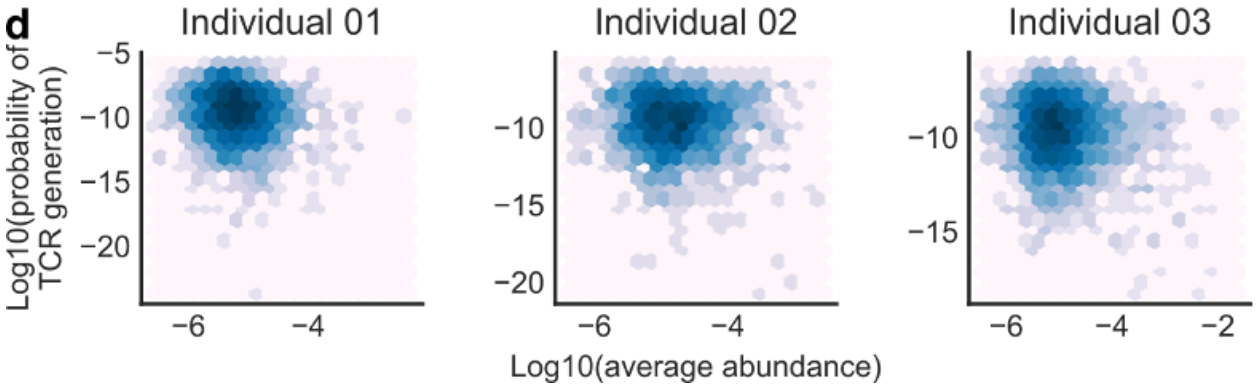


Figure S15. Persistent and public receptors may result in part from TCR recombination biases. (**a**) As in previous studies, the probability that a given TCRβ was generated correlated closely with publicness in a cohort of 778 individuals. For each individual, only TCRβs occurring in both that individual and the cohort were considered. (number of TCRβs evaluated in individual 01 = 638091, Spearman *rho* = *0.51111,* $p < 10^{-6}$; number of TCRβs evaluated in individual 02 = 338617, Spearman *rho* = 0.52231, $p < 10^{-6}$; number of TCRβs evaluated in individual 03 = 284990, Spearman *rho* = 0.51129, $p < 10^{-6}$). (**b**) TCRβs occurring at more time points tended to have higher generation probabilities, although persistent TCRβs did not have higher generation probabilities than other receptors observed at multiple time points. Letters indicate significant differences from all other groups by a Mann-Whitney *U* test ($p < 0.001$), while dagger (†) indicates groups that were not significantly different from multiple other groups. (**c**) Mean abundance of all TCRβs correlated significantly with generation probability but with a low correlation coefficient (individual 01: Spearman *rho* = 0.07884, $p < 10^{-6}$; individual 02: Spearman *rho* = *0.05300,* $p < 10^{-6}$; individual 03: Spearman *rho* = 0.08208, $p < 10^{-6}$). (**d**) Mean abundance of persistent TCRβs did not correlate with generation probability (persistent TCRβs: number of TCRβs evaluated in individual 01 = 3448, Spearman *rho*=-0.08988, $p < 10^{-6}$; number of TCRβs evaluated in individual 02 = 1978, Spearman *rho* = -0.04341, $p = 0.0537$; number of TCRβs evaluated in individual 03 = 2965, Spearman *rho* **=** 0.04552, $p = 0.01318$).

Table S1 (available online). Overall TCRβ-sequencing statistics per sample: sequencing depth, productive TCRβ sequencing depth, fraction of productive TCRβ sequences, unique V genes identified, unique J genes identified, unique CDR3 sequences, unique TCRβs, unique TCRβ nucleotide sequences.

Table S2 (available online). V gene usage across subject and T cell population, expressed as both a fraction of all unique productive TCRβs and as a mean total abundance per sample.

Table S3 (available online). J gene usage across subject and T cell population, expressed as both a fraction of all unique productive TCRβs and as a mean total abundance per sample.

Table S4 (available online). Sequence and abundance information for the largest cohort of closely correlated TCRβs identified in eacxh individual by Spearman's or Pearson's correlation.

Table S5 (available online). The fraction of TCRβs in each sample that occurred in 1-8 samples from that subject's time series.

Table S6 (available online). Mann-Whitney *U* test statistics for mean abundance of TCRβs occurring in different numbers of samples during the time series.

Table S7 (available online). Mann-Whitney *U* test statistics for nucleotide redundancy of TCRβs occurring in different numbers of samples during the time series.

Table S8 (available online). The fraction of TCRβs in each sample that were shared to different degrees among subjects in a large, independent cohort.

Additional File 1 (available online). Count, frequency, V, J, and CDR3 amino acid sequence data for high abundance TCRβs in each sample. This is a tab-delimited, gzip-compressed file. The column "tcr" corresponds to a label that identifies a unique combination of CDR3 amino acid sequence, V gene, and J gene observed in this study. The column "count (templates/reads)" represents the counts of a given receptor's DNA sequence in the sequencing data. The column "frequencyCount (%)" is the relative abundance of that TCR within each sample, accounting for productive and nonproductive receptors. The columns "vGeneName" and "jGeneName" are the IMGT assigned V and J genes. The column "aminoAcid" is the CDR3 amino acid sequence.

Additional File 2 (available online). Data for persistence across our time series and sharing across subjects in an independent, large cohort for TCRβs in this study. The columns "aminoAcid", 'vGeneName", "jGeneName", and "tcr" are the same as for Additional File 1. The column "n-cmv-public" is the number of subjects (out of 778) that shared that TCRβ. The columns "numocc_sublpbmc", "numoccsub2_pbmc", and "num_occ sub3_pbmc" are the number of time points at which a given receptor was observed in the PBMC samples for Individual 01, 02, and 03, respectively.

Conclusion

The common thread through the three results chapters in this thesis is that they demonstrate the value of greater sampling to achieve finer resolution of study observations and to discover hidden heterogeneity. We showed this as applied to both longitudinal and cross-sectional studies. Here I will reflect on the potential significance of our results and discuss some limitations and potential extensions for our work.

Chapter 1

In chapter 1, we reported the longitudinal tracking of gut microbiomes of a small cohort of travelers over a long period with frequent sampling. With this kind of study it is hard to standardize the conditions in which we would like to study subjects since there are many variables at play, such as duration of stay, regional destination, and behavioural differences while abroad. Short of planning subjects' travel for them (and what kind of research budget would allow that), the best we can do is to be transparent with intra-study variability and potential confounders.

The tradeoff we took in our study for participants who were compliant over time and trained in self-sampling was that there were few of them who fit this bill, which in some ways is the opposite tradeoff of most travel studies in the literature (Cheung et al. 2023; D'Souza et al. 2021; Kampmann et al. 2021). This carries implications for how our results should be interpreted, and to what extent we can reasonably expect results to generalize.

Some of our results, particularly those involving more minor observed changes or inconsistency among even our small cohort, are less likely to hold water upon more rigorous validation. We reported these to be thorough with our analysis. On the other hand, several of our findings appear to corroborate each other, chiefly our recurring observation of the lack of destination-specific effects and local microbial uptake. These present interesting leads to pursue for future work aiming to expand and solidify our understanding of travel and the microbiome.

There are a few natural follow-up directions for our work. The most obvious (but at the same time very challenging) is to expand the cohort size, perhaps by following a group of travelers at once, which may also better ensure uniformity in travel duration, origin, and destination (with the tradeoff of less “independent” observations if travelers travel together and possibly less generalizable). As well, since we employed amplicon sequencing for our samples, which is more economical but provides less rich information, another next step could be to instead use shotgun metagenomic sequencing, which can not only identify which bacteria are in a sample, but also give readouts on bacterial genome content (and thus inferred functional capacity, including antimicrobial resistance). The resulting higher-dimensional data would be interesting but also a greater task to analyze.

Chapter 2

The results we presented in chapter 2 constitute a slice of all preliminary analyses to accompany the introduction of this immense resource. Limitations to our results from the analysis side chiefly stem from our reliance on existing databases comprising already-identified systems, and making assumptions about how they extrapolate. From the wet lab side, the questions of donor diversity and how representative they are of global microbiome diversity are pertinent.

The potential follow-ups are endless, both computational and experimental (since this resource comprises not only a rich dataset but also a physical strain library). Our aim is to encourage further exploration, leveraging our resource.

Chapter 3

Many of the study design limitations and extensions for chapter 3 mirror those for chapter 1, as both are longitudinal profiles of small subject cohorts. I will elaborate on some points of discussion specific to chapter 3.

First, while the TCRβ CDR3 is a good choice for single-locus profiling when studying TCR repertoires, it is not the sole determinant of antigen specificity. This is true on a few levels. Within the TCRβ chain, there is more to the V region than just the CDR3, and it

is possible that our choice of sequencing bounds was conservative. The TCRβ chain is also one half of a two-chain complex with the TCRα chain, which also contacts MHC-antigen. Beyond this, while α/β T cells are a good choice to prioritize when profiling TCR repertoires, they do not comprise the entirety of the T cell population in the body. To tie things back in with the gut microbiome, γ/δ T cells, for example, while a minority of all T cells, are known to be highly involved in the gut-associated lymphoid tissue (Wu et al. 2023). Thus, expanding the sequencing target, for example by capturing the whole V region, or capturing paired α/β chain information (such as by single-cell protocols which now exist) (Redmond, Poran, and Elemento 2016), or even capturing more T cell types, could provide additional layers of information beyond what we had in our study.

There may also be value in further separating by cell sublineage, whether physically by sorting using T cell surface markers, or computationally by examining gene expression. It could also be interesting to sample not only blood, but different peripheral tissues, where certain T cells may localize. This could provide a more in-depth picture of immune activity of these different cell types.

Finally, our design was to sample (up to) monthly over a period of one year, which is likely relevant for certain known patterns within the immune system, such as seasonal exposures. There may, however, be finer-scale dynamics, such as on daily timescales, that average out in our data. Increasing sampling frequency (again, mirroring what we did in chapter 1) would capture these finer-grained processes. It may also be interesting to time the onset of sampling with the onset of an anticipated immune response. Most recently, this has been employed in time course studies of repertoires in immune responses to COVID-19 (Kotagiri et al. 2022).

References

Cheung, Man Kit, Rita W. Y. Ng, Christopher K. C. Lai, Chendi Zhu, Eva T. K. Au, Jennifer W. K. Yau, Carmen Li, et al. 2023. “Alterations in Faecal Microbiome and Resistome in Chinese International Travellers: A Metagenomic Analysis.” *Journal of Travel Medicine* 30 (6). https://doi.org/10.1093/jtm/taad027.

D’Souza, Alaric W., Manish Boolchandani, Sanket Patel, Gianluca Galazzo, Jarne M.

van Hattem, Maris S. Arcilla, Damian C. Melles, et al. 2021. “Destination Shapes Antibiotic Resistance Gene Acquisitions, Abundance Increases, and Diversity Changes in Dutch Travelers.” *Genome Medicine* 13 (1): 79.
Kampmann, Christian, Johan Dicksved, Lars Engstrand, and Hilpi Rautelin. 2021. “Changes to Human Faecal Microbiota after International Travel.” *Travel Medicine and Infectious Disease* 44 (November): 102199.
Kotagiri, Prasanti, Federica Mescia, William M. Rae, Laura Bergamaschi, Zewen K. Tuong, Lorinda Turner, Kelvin Hunter, et al. 2022. “B Cell Receptor Repertoire Kinetics after SARS-CoV-2 Infection and Vaccination.” *Cell Reports* 38 (7): 110393.
Redmond, David, Asaf Poran, and Olivier Elemento. 2016. “Single-Cell TCRseq: Paired Recovery of Entire T-Cell Alpha and Beta Chain Transcripts in T-Cell Receptors from Single-Cell RNAseq.” *Genome Medicine* 8 (1): 80.
Wu, Xiaoli, Jun Yan, Zhinan Yin, and Guangchao Cao. 2023. *γδ T Cells in Physiology and Pathology*. Frontiers Media SA.