מכון ויצמן למדע
WEIZMANN INSTITUTE OF SCIENCE

Thesis for the degree
Master of Science

Submitted to the Scientific Council of the
Weizmann Institute of Science
Rehovot, Israel

עבודת גמר (תזה) לתואר
מוסמך למדעים

מוגשת למועצה המדעית של
מכון ויצמן למדע
רחובות, ישראל

By
Liron Zahavi

מאת
לירון זהבי

שונות גנומית תוך-מינית במיקרוביום מעי האדם והקשר בינה לבין תכונות המארח

Intra-species genomic variability in the human gut microbiome and its association with host factors

Advisor:
Prof. Yitzhak Pilpel

מנחה:
פרופסור יצחק פלפל

April 2018

ניסן תשע"ח

Abstract

A growing body of evidence supports associations between the human gut microbiome and the health of its host, yet the mechanisms underlying this interaction are still not fully understood. Typical microbiome studies focus on species composition, however modern technology allows analyses at the sub-species level. Recent studies demonstrated that genetic diversity at the sub-species level of the microbiome expresses selective forces acting upon bacterial genomes in the gut, and may promote our understanding of mechanisms underlying host-microbiome interaction and bacterial adaptation. In this project, we devised a computational framework to analyze intra-species genomic variability in the human gut microbiome. We have applied this framework to dozens of bacterial species across a cohort of 1003 human participants. We compared the genetic variability along the genome of *Akkermansia muciniphila* and found a difference in variability patterns between genes of different functions, such as different variability of protein coding genes and non-protein coding genes. We further compared the genomic variability of different symbiont species and different hosts and integrated extensive information about the host into the analysis. To our knowledge, this is the first analysis of single-nucleotide-resolution genomic variability landscape of the human gut microbiome in association to host phenotypes. This revealed associations between genomic variability in the microbiome and host health and lifestyle factors. We found, for example, that the microbiome of smoking hosts, and the microbiome of hosts with high white blood cells counts, display higher intra-species genomic variability. These results demonstrate how analyzing intra-species variability in the gut microbiome of different individuals may shed light on host-microbiome interaction and on bacterial genetics.

Table of contents

List of abbreviations

SNP - single nucleotide polymorphism

WBC - white blood cells count

Introduction

While every species has a typical genomic sequence, individuals in the population will not have entirely identical genomes. Genetic variation exists in every biological unit: between species, within populations of same species (Altshuler et al., 2010), and even among cells of a multicellular organism (Martincorena et al., 2015). The source of the variation is mutation, a random error in the genome replication, which given typical error rates of DNA replication enzymes and genome lengths of bacteria, for example, is likely to occur once in hundreds of cell divisions thus always maintaining some amount of genetic variation (Zhao et al., 2017). When a mutation occurs, it may have a deleterious influence on the individual's fitness, it can be beneficial, or, it will be a neutral mutation, which has no influence on the individual's fitness. In a genetically diverse population, some individuals will have a genetic variation which gives them a fitness advantage in their environment. Under some selective pressure they are more likely to survive, making their genetic variants more abundant in the next generation and the population better adapted to its environment. Genetic variability is the basis for evolution (García-Arenal, Fraile, & Malpica, 1999) and was essential for the emergence of every species that has ever existed.

The human gut is populated by many microorganisms, most of which are bacteria, often referred to as the 'microbiome' or 'microbiota'. Recent quantification estimates the number of bacterial cells in our gut to be as high as 10^{13} (Sender, Fuchs, & Milo, 2016), as many cells as there are in the human body. This bacterial population closely interacts with its human host, providing important functions while relying on it for their survival. These bacteria have been shown to have great influence on the host health, associated with various health conditions such as obesity (Turnbaugh et al., 2009), diabetes (Larsen et al., 2010), inflammatory bowel disease (Frank et al., 2007), drug metabolism (Clayton, Baker, Lindon, Everett, & Nicholson, 2009) and even with neurodevelopmental disorders (Hsiao et al., 2013). The mechanisms of interaction between the symbiont bacteria and the health of their human host are far from being fully understood, however their importance is demonstrated over and over in the published scientific literature.

Not only does the bacterial population in our gut influences our health, our health and lifestyle are affecting the gut niche and thus shape the population of gut microbiota. The influence of host lifestyle on the gut residing bacteria population was observed in multiple levels of variation, including species composition (David et al., 2015) and mobile genes distribution (Brito et al., 2017). In their study, Brito et al. have described the acquisition of different mobile genes by the microbiome bacteria, in association with the host diet and antibiotics consumption. Studying the mechanisms through which bacteria adapt to its host's gut is key to our understanding of host-microbiota interactions as well as of microbiome stability, which may be essential for designing effective microbiome-based therapies.

As the human host's physiology and habits change the conditions in the gut environment, the genetic variability of the residing bacteria allows its adaptation. Another important level of genetic variability in the gut is the sequence diversity of different nucleotides in the genome. As more than a billion point mutations are expected to occur in our gut microbiome every day (Lieberman, 2018), the potential of the microbial population to evolve and adapt to its host is great. Zhao et al. have recently described the evolution of a single bacterial species in the human gut and revealed within-host evolution, expressed by its diversity of mobile genes and single nucleotide polymorphisms (SNPs) (Zhao et al., 2017). In their analysis, the authors have found that certain genes contain significantly more SNPs, implying on their relevance for niche adaptation. Comparing the genetic variability of different genes may imply on the function and essentiality of the gene: a gene that may be important for bacterial survival in a given niche may be significantly conserved, while a gene which is under strong positive selection to adapt to the environment may shed light on the environment sensed by the bacteria and on host-bacteria interactions. Schloissnig et at. have analyzed SNPs in a variety of gut microbiota species and in multiple hosts, and by comparing selection marks on genes of two commensal species, have showed how stronger purifying selection on a metabolic gene in one of the species correlates with previous knowledge of the differences between the metabolism of these two species (Schloissnig et al., 2013). These studies imply that sub species-level genetic diversity in the microbiome expresses the selective forces acting upon bacterial genomes in the gut and demonstrate its potential to highlight specific genes and mechanisms through which bacteria adapt.

While both studies describe variation between hosts in the amount and distribution of microbial genetic variability in the microbiome, neither have compared these differences with different phenotypes or habits of the host. Zhao et al. found several genes to be under subject-specific selective forces, and several others to be under selective pressures common to all hosts in the cohort. Integrating this sort of observations with information about the human host may shed light on human lifestyle habits and medical conditions which challenge the commensal bacteria, and the mechanisms through which the latter adapt. Analyzing the genomic variability in the context of the human host, may promote our understanding of how the variability is induced by environmental features, as well as of how it affects the host health.

In this study, we have devised a computational framework to analyze intra-species genomic variability on a single nucleotide resolution, and used it to analyze the genomic variability of dozens of bacterial species in the microbiome of hundreds of human hosts. We have described the genetic variability landscape in the genome of one bacterial species across the entire cohort, comparing different regions and genes in the genome. Using an ample dataset of microbiome samples and extensive host medical and lifestyle information, we have compared the genomic variability of many species, in many hosts, and tested its association with host health and lifestyle.

Goals

We have set the following goals for this project:

- Create a computational framework to analyze intra-species genomic variability in bacterial populations, on a single nucleotide resolution, based on sequenced metagenomic samples.
- Use this framework to map the *in-vivo* genomic variability landscape of many bacterial species from divergent gut microbiome populations in a systematic manner.
- Learn about bacterial genetics by quantifying the intra-species genetic diversity along bacterial genomes and searching for regions with extreme variability or conservation.
- Combine the genomic variability landscape of the microbiome with detailed data regarding its human hosts to shed light on genetic and environmental factors associated with variability. Analyze the evolutionary forces acting upon bacterial genomes to learn about mechanisms of bacterial adaptation to the host gut environment.

We expected these goals to promote our understanding of bacterial evolution and host-microbiome interactions.

Methods

Gut microbiome samples and human hosts data

Data for this project were taken from multiple cohort studies published by the Prof. Eran Segal and Prof. Eran Elinav groups (Korem et al., 2017; Zeevi et al., 2015) including gut microbiome samples and human host data from 1003 healthy participants aged 18-70 years. Participant data includes medical and lifestyle information collected by blood tests, anthropometric measures, heart rate and blood pressure measurements, medical history, lifestyle and food frequency questionnaires.

Microbiome data are in the form of metagenomic reads collected from stool samples and sequenced as detailed in the published papers (Korem et al., 2017; Zeevi et al., 2015). For some hosts, multiple metagenomic samples were included, either sampled from the same stool (using two different collection kits) or at different time points (which can be days to years apart).

Mapping metagenomic samples to bacterial genomes

Single end reads were mapped to bacterial reference genomes to search for the genome and the position in the genome from which this sequence originated. Mapping was done using GEM (Marco-Sola, Sammeth, Guigó, & Ribeca, 2012), based on sequence similarity. Reads which were assigned to the human genome were excluded. Mapping parameters used are same as in Zeevi et al.,2015.

In the first part of the project, we measured the genetic variability along the genome of *Akkermansia muciniphila* over the entire cohort by mapping all metagenomic samples to NCBI genome of *A. muciniphila,* ATCC BAA-835 (Agarwala et al., 2017). Only reads mapped uniquely to this species were kept.

In the second part of the project, we analyzed all the bacterial species present in the cohort in a more systematic manner. We mapped the reads from the metagenomic samples to proGenomes dataset (Mende et al., 2017) of bacterial genomes. The results of the mapping algorithm were further analyzed by an iterative algorithm for mapping corrections based on PathoScope (Hong et al., 2014) which assigns probabilities to different mapping destinations based on species abundance. For downstream analyses, we only used reads with a genome assignment with probability equal to or greater than 0.99 (Results).

Creating nucleotide occurrences array for each genome

We used the assignment of reads from the metagenomic samples, indicating from which position in which genome it originated, to create a matrix summing the number of observations of each nucleotide (‘A’, ‘T’, ‘C’, ‘G’, or ‘-’ - representing a short deletion in comparison to the reference genome) in each genome position. We represented each species’ genome in each sample by a matrix of size [genome-length x 5] (5 – for the 4 nucleotide types, plus a indel). We later removed the indel column for separate analyses. This resulted in a

nucleotide occurrences array with 4 columns and N rows, with N being the length of the bacterial genome for each species in each sample.

Creating nucleotide frequency array

We normalized each row (representing a position in the genome) in the nucleotide occurrences array by dividing each value in the row (representing the number of occurrences one of the nucleotides) by the sum of the row (representing the total number of reads which were mapped to this position).

Creating *A. muciniphila* nucleotide frequency array

For the analysis of *A. muciniphila* genome, we were interested in quantifying genetic variability in different genomic regions across the entire cohort. To get as much information as possible, we selected all metagenomic samples with an average genomic coverage of 0.5X or more for *A. muciniphila* and created the nucleotide-frequency array described above for each sample. Two analyses, the CRISPR array detection (Results) and the comparison of genes' variability distribution within individual samples, were done only with the 0.5X coverage threshold. For the rest of the analyses of the *A. muciniphila* genomic variability we have added a threshold and for a coverage by 6 or more reads for a genomic position to be included in the analyses.

After creating a nucleotide frequency array for each sample that compiled with these thresholds, we summed all sample-arrays to create one merged *A. muciniphila* array.

Finally, we renormalized each position of the merged array, resulting in an array of nucleotide frequencies for each position of *A. muciniphila* genome, over the entire sampled cohort. The reason for normalizing the array twice (once for each sample separately, and once again after summing all sample arrays), rather than just summing all sample arrays and only then normalizing the array, was to conform equal weights in the analysis for high-coverage and low-coverage samples alike.

Creating nucleotide frequency array for all species in all samples

In the second part of the project, we analyzed the genomic variability of each species within each sample separately. For several analyses (Results), to increase coverage, we merged all reads from all samples of the same host into one unified sample to increase coverage (prior to any frequency calculation or normalization step). However, to prevent intra-host bias resulting from sampling across multiple time points, we selected one sample from each participant when comparing variability across different hosts. In these analyses, we tried to eliminate as many of the technical differences between samples as possible. Therefore, we chose the samples which were all collected using the same stool collection kit and sequenced on the same sequencing platform. Since all samples were down-sampled to contain no more than 10^7 reads, post human-DNA-filtering, we also excluded all samples with less than 10^7 reads, to have an equal number of reads in all samples.

Quantifying genetic variability

To quantify the variability in each position, based on the nucleotide distribution, we calculated its entropy (Shannon, 1948):

$$entropy = \sum_{n \in \{A,T,C,G\}} P_N(n) \times \log_2[P_N(n)]$$

$$P_N(n) \coloneqq \textit{the relative abundance of nucleotide } \boldsymbol{n} \textit{ in this position}$$

There are many different measures commonly used to quantify sequence variability, such as 'major allele frequency', or the relative abundance of the most abundant allele for each position, and 'minor allele frequency', or is the relative abundance of the second most abundant allele in the position. We chose to use entropy, since it contains information both on the number of different nucleotides observed in the position, and on the distribution of reads across different possible nucleotides. Theoretically, for DNA sequences entropy values can range from 0, for a position where the entire sampled population has the same nucleotide, to 2 for a position where all four possible nucleotides are equally abundant.

We calculated the entropy of each position of a genome-array separately, creating an entropy vector (with the length of the genome) for each species' genome in each host/sample. This enabled us to compare the variability of different genomic regions within a sample, as well as the variability of a certain genetic unit in one sample in comparison to the equivalent genetic unit in another sample.

Calculating average genomic entropy

Taking the nucleotide-frequency array of a genome, we calculated the entropy of each position covered by five reads or more. Next, we discarded genomes that had fewer than 10,000 such positions. In each of the arrays which passed this filtering, we calculated its average entropy over all covered positions.

Samples and species selection

After mapping all samples, we examined the list of reference genomes which had reads assigned to, to search for multiple strains of the same species having reads mapped to them, and excluded these species entirely from further analysis. This was done since we only wanted to analyze the intra-species variability, and because we did not know how well the sequencing pipeline distinguishes between different strains of the same species.

Calculating sample entropy score

To score the overall intra-species variability of all species in a sample, we gave it an entropy score. For each species in the sample, we calculated a Z-score quantifying how much its average-genomic-entropy in this

sample deviates from the average value for this species. Next, we calculated the average of the Z-scores of all the species in the sample, which we describe as the sample entropy score.

Calculating species entropy score

To score the overall variability of a species, calculated over all its sampled populations, we gave it an entropy score as well (similarly to the mentioned above 'sample entropy score'). This time, for each sample which included this species, we calculated a Z-score for the sample, quantifying how much the average-genomic-entropy of this species deviates from the average value for this sample. Then, we calculated the species entropy score as the average of all Z-scores of samples which contained this species.

A. muciniphila total coverage analyses

After we created the nucleotide occurrences array (counts the different nucleotides for each position) for *A. muciniphila* in each of the samples and filtered out low coverage samples (same filters as in 'Creating *A. muciniphila* nucleotide frequency array'), we summed all occurrences arrays together, resulting in one array summarizing the number of reads containing each of the 4 nucleotides, for each genomic position, in all of the samples together. Next, we summed all the nucleotides' occurrences in each position- summing each row, which created a vector with the length of *A. muciniphila* genome, counting the total number of reads (from all samples together) that mapped to each genomic position, i.e. the total coverage of each position.

Calculation of average entropy per codon position

For every protein-coding gene in each genome (each bacterial species in each sample), we calculated separately the entropy of the first, second and third codon positions, using the genome-wise entropy vector. Next, for each such genome, we calculated the average entropy of the first, second and third codon positions.

In each genome we only calculated the entropy in positions covered by 4 reads or more, and only genes which at least one third of their positions were covered. Since some genomes had sufficient coverage for only a gene or two, for each species we kept only samples in which the number of covered genes was not lower than the median number for this species, and not smaller than 15 genes.

Calculation of the ratio between the first two codon positions and the third

After calculating the average entropy of each codon position of a genome as described above, we calculated the ratio between the positions, using this formula:

$$ratio = \frac{\frac{1}{2} \times (1^{st} positions\ mean\ entropy + 2^{nd} positions\ mean\ entropy)}{3^{rd}\ positions\ mean\ entropy}$$

Where either the numerator or the denominator were zero, we replaced them with the lowest non-zero value (lowest numerator or denominator, accordingly) calculated in all species and samples.

Alpha diversity calculation

We calculated the alpha-diversity of a sample by calculating the average genomic coverage of each bacterial species in the sample, then using them to calculate the diversity index based on Shannon entropy (Shannon, 1948):

$$Alpha\ diversity = \sum_{s \in \{species\}} P_S(s) \times \log_2[P_S(s)]$$

$$P_S(s) \coloneqq the\ relative\ abundance\ of\ species\ \boldsymbol{s}\ \ in\ this\ sample$$

Results

Genomic entropy landscape of *Akkermansia muciniphila*

As a proof of concept for the importance of genomic variability in the human microbiome, we first explored the prevalence and scale of it in the human symbiont *A. muciniphila*. As a first step, we calculated the nucleotide distribution in each position of the *A. muciniphila* genome in each human microbiome sample. We next calculated the per-position entropy over the entire cohort, as a proxy for the overall variability in each genomic position (Methods).

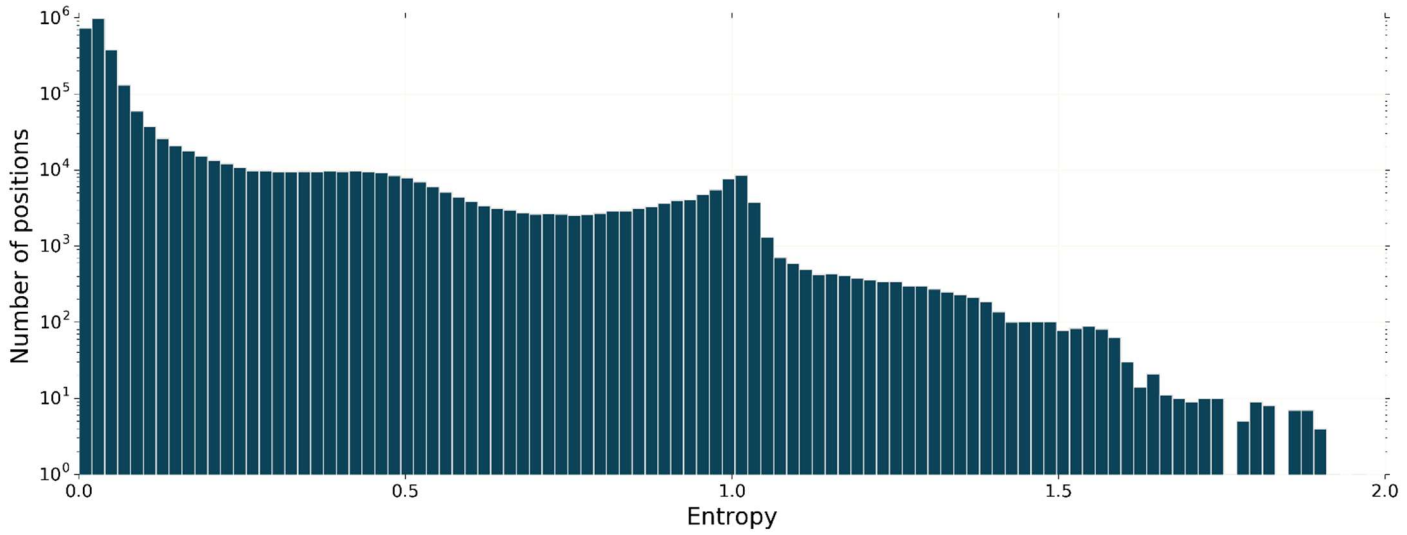

Figure 1 – **per-nucleotide entropy distribution in the *A. muciniphila* genome is non-uniform.**
Theoretical entropy range is 0, for no variability, to 2, for the highest variability. If all reads mapped to a position agree on the same nucleotide, the entropy is zero. In a genomic position where two nucleotides are equally abundant, the entropy is 1. In a position where all four nucleotides are equally abundant the entropy is 2.

Comparing the variability of different positions along the genome (Fig. 1) we show that although positions which are identical across all sampled genomes are not highly abundant – we observed zero entropy in only 3.38% of the positions, 95.49% of the positions had entropy values between 0 and 0.5, which is the lower quarter of the possible values range, signifying considerable conservation in most genomic positions.

To increase the accuracy of our method, we filtered regions of aberrant coverage from our calculations. To this end, we used metagenomics data to calculate read coverage in different regions of the genome and across different samples. Briefly, we summed the number of reads mapped to each genomic position in each sample (Methods). Our examination of read coverage of the *A. muciniphila* genome across the entire cohort, revealed highly variable coverage (Fig. 2), whereby several regions of the genome are covered up to 4.5-fold more than the average coverage, and other regions have zero coverage in all samples. Such differences in coverage may result from differences in gene copy number between the reference genome and the genomes of bacteria present in samples. Another possible cause for variability in coverage is sequence homology of a

region with other species. Genomic regions presenting significantly higher-than-average coverage may be result from the presence of a different, highly abundant species in the microbiome harboring a homologous sequence in its genome which is not represented in the reference database and is therefore associated with *A. muciniphila*. Similarly, regions with significantly low coverage may result from the 'sharing' of metagenomics reads between two or more species. This demonstrates the importance of correct mapping to separate intra-species variability from inter-species variability. Thus, to improve our ability to separate the former from the latter, we have excluded reads with ambiguous mapping, and for some of the downstream analyses we have also set a threshold of minimal genomic coverage for inclusion of genomes (Methods).

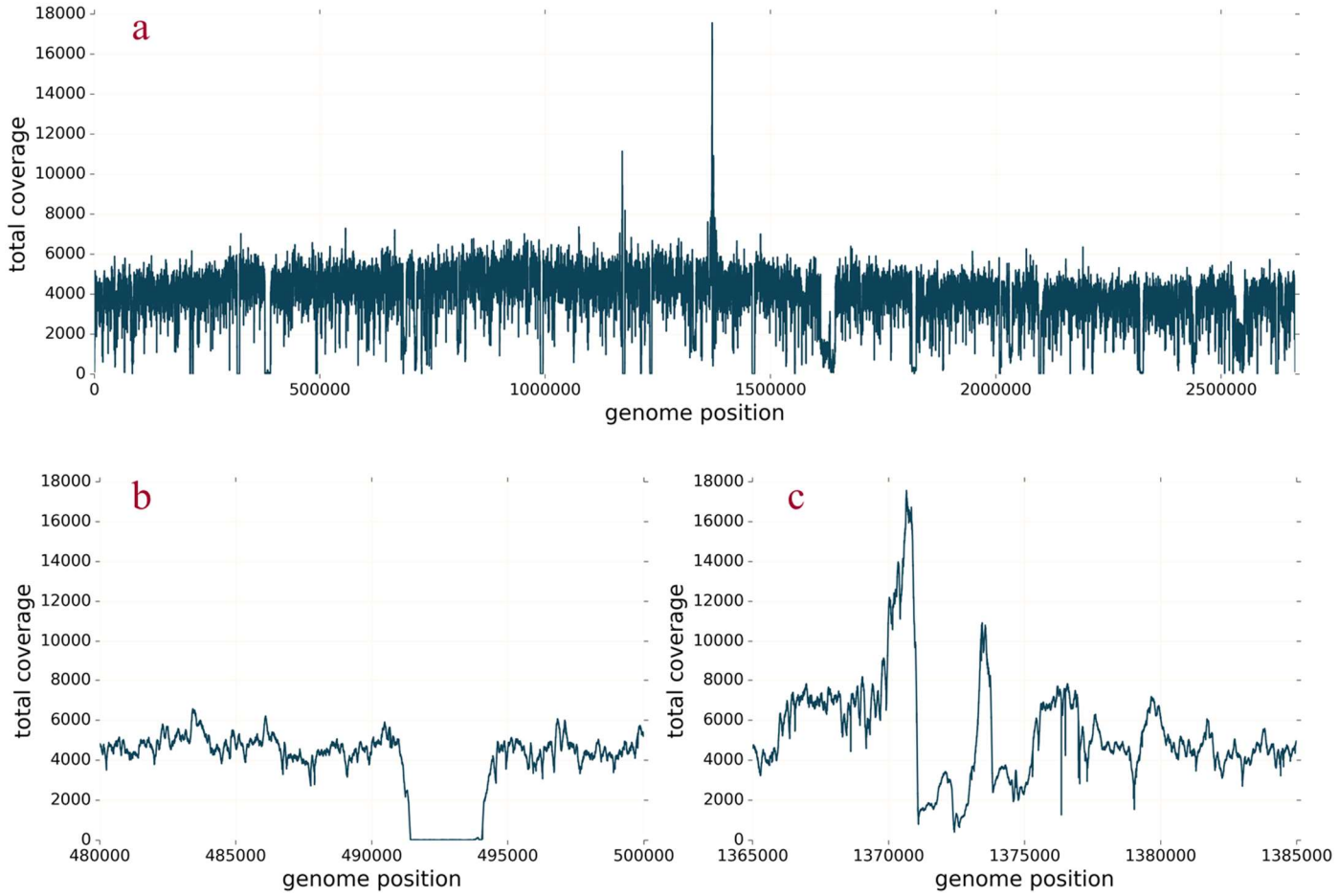

Figure 2 – **High variability in the coverage of different regions of *A. muciniphila* genome**.
Number of reads mapped to each position of *A. muciniphila* genome. **a.** all positions of *A. muciniphila* genome. **b.** an example for a region with zero coverage. **c.** an example for a region with highly variable coverage

Genetic variability of different *A. muciniphila* genes

Genes are the basic unit of function in all life kingdoms and are key to our understanding of bacterial function and adaptation. To advance our understanding of the relation between genomic variability and function, we compared the variability across different *A. muciniphila* genes.

Calculating the per-position entropy distribution in all gene positions, we observed that while most of the positions in a gene have almost no variability (represented by the narrow distribution of gene's median entropy around zero)- the mean entropy in genes have a wider range of values (Fig. 3).

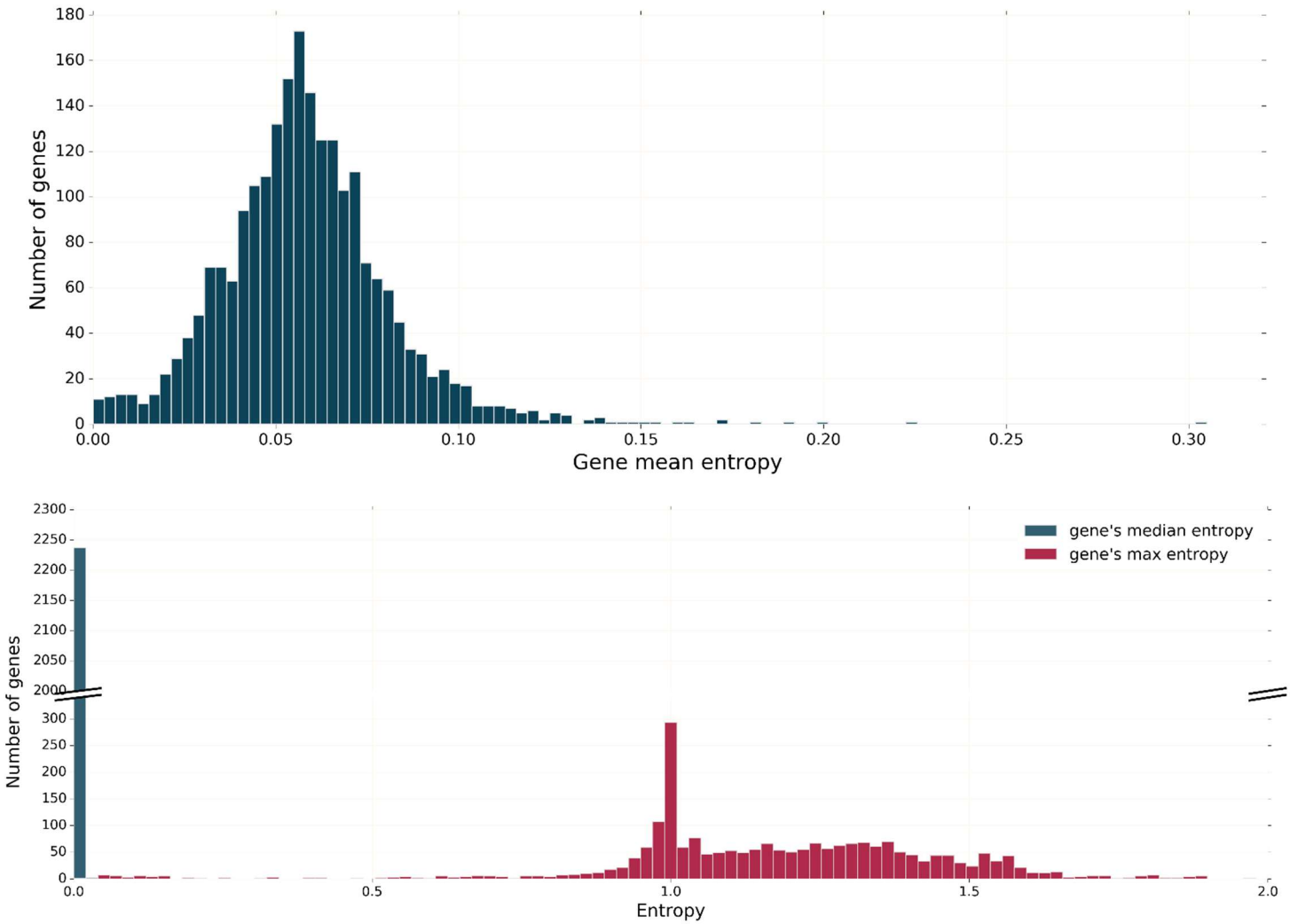

Figure 3- **Variability of different genes in *A. muciniphila* genome is different** a. distribution of mean gene entropy. Mean gene entropy is calculated as the average of entropy values in all gene positions. b. distribution of median gene entropy and maximal gene entropy.

We found the 16 least variable genes, and 33 out of the 50 least variable genes, to be transfer-RNA genes. This further supports the link between reduced variability, or reduced entropy and the importance of function, as these genes which code for instrumental translation machinery are highly essential for cellular function. As these genes are not translated to proteins, they do not have the sequence flexibility that the genetic code normally allows protein-coding genes to have, and thus mutations in these genes are expected to have a greater fitness impact on the individual, making them less likely to spread in the population. Three additional RNA genes were found among the 50 least variable genes, all expressing 5S ribosomal RNAs. Seven of the 50 least variable genes are labeled as 'hypothetical proteins', which were not yet annotated. More than a third (900 of 2374) of *A. muciniphila* genes are labeled as 'hypothetical proteins', and the relative stability of these seven

genes can imply on their significance and highlight them for further analysis. Their position on the variability spectrum, which is similar to that of the non-coding RNA genes mentioned above, suggests that these unannotated genes may also express functional RNA. When we compared the entropy distribution in protein coding genes, excluding those which are labeled as 'hypothetical proteins', with the entropy distribution in RNA genes, we found the entropy to be significantly lower for genes which are known to code for functional RNAs (Mann-Whitney $p < 10^{-29}$).

Most of the genes harboring the highest variability are annotated as coding for hypothetical proteins. Other sequences within the top 50 most variable genes are of unknown functions, however they contain domains which are known to be associated with membrane proteins and transporters. In the process of genome evolution, inessential genes often accumulate mutations due to the relaxation of selective pressures, and duplicated genes often speciate to obtain new functions. Such highly variable genes may belong to either of these categories, with membrane and transporter genes possibly evolving as sensing and response mechanisms for environmental, immune or other cues. As in this analysis we have measured the total genetic diversity of these genes in the cohort, the source of this diversity can be intra-populations (i.e. intra-sample) variability, inter-populations variability, or a combination of both. We suggest future studies to examine whether these genes have a small intra-populations variability and high inter-populations variability, suggesting an adaptation mechanism to specific hosts through these genes.

We have also calculated the intra-sample variability of different genes for the five samples with highest coverage of *A. muciniphila*. We found that the intra-sample, or intra-population, variability is also different for different genes (Fig. 4), fortifying our motivation to compare the variability of genomes in different samples in future analyses.

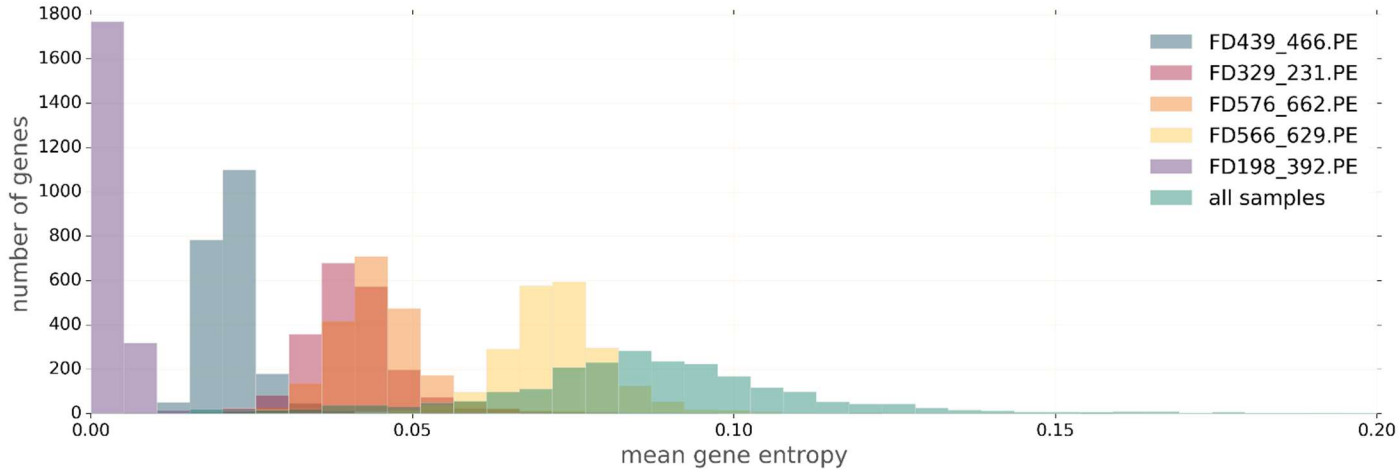

Figure 4 – **Gene entropy distribution varies between different samples**

We further explored genomic regions with extremely high variability, and observed a pattern resembling the spacers within a CRISPR array (Fig. 5). Using existing gene annotations for this genome (Agarwala et al.,

2017), we found that a known repeat region is located in this region, and that the two neighboring genes are known CAS genes. Thus, we analyzed the consensus sequence of this region (choosing for each position the most abundant nucleotide in the population) using a CRISPR finding tool (Grissa, Vergnaud, & Pourcel, 2007) and confirmed that this region is indeed a CRISPR array, with the low entropy regions being the repeat sequences and the high entropy regions being the spacers. This is with agreement with previous studies of CRISPR arrays (Sorek, Kunin, & Hugenholtz, 2008; Tyson & Banfield, 2008), which reported high variability in the spacer sequences. This suggests that looking for specific variation patterns with this data might reveal new CRISPR arrays in unannotated bacterial genomes, as well as additional unknown variability structures pertaining to novel regulatory and functional regions.

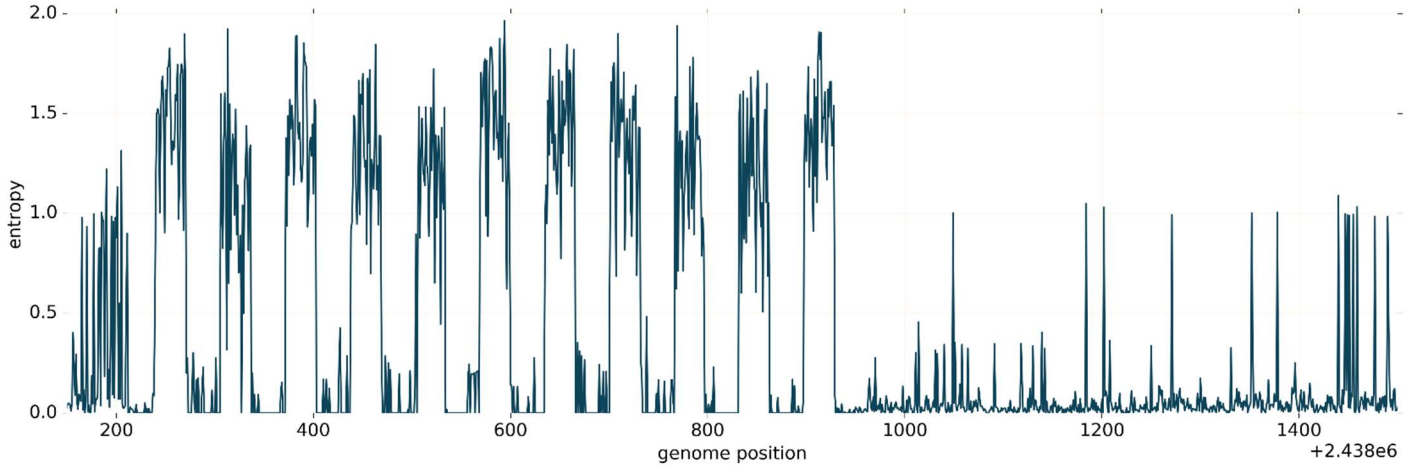

Figure 5- **Per-position entropy along a CRISPR array reveals a unique pattern** peaks and valleys of entropy mark the positions of spacers and repeat sequences, accordingly. Two neighboring genes (right side of the figure) are coding for CAS proteins.

Genetic variability of different positions within a gene

We next compared the entropy of different positions along each gene and revealed a three-nucleotide periodicity in protein coding genes, reflecting the degeneracy and degrees of freedom in the genetic code (Fig. 6, top). In each codon, the third nucleotide is more variable on average than the previous two, as explained by the wobble effect of the third codon position (Crick, 1966).

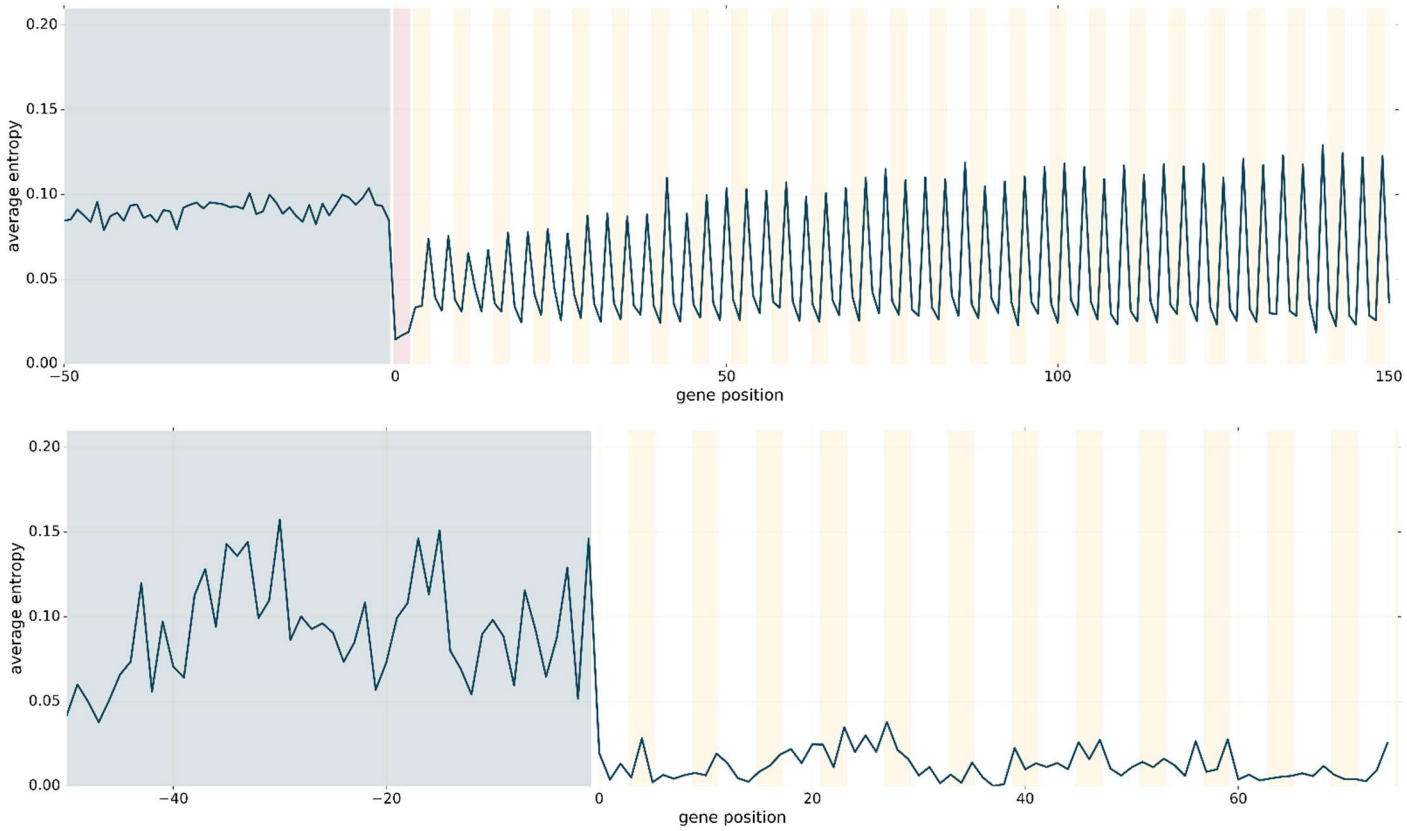

Figure 6 – **Variability of protein coding genes reveals a periodic pattern** per-position entropy, averaged across all genes in *A. muciniphila* genome. X axis zero marks the position of the first nucleotide of the gene. The area colored in grey marks the 50 nucleotides upstream to gene start position. Every three nucleotides are colored in yellow and white, alternately. In protein coding genes, the first 150 first positions of the gene were analyzed, while in non-protein coding genes only the first 74 positions of the gene were analyzed, in order to be able to include short genes such as tRNAs. (top) average entropy in each position of protein-coding genes, averaged over 2254 genes. Grey part is the 5'-UTR, and the start codon region is marked in red. (bottom) non-protein coding genes, averaged over 60 genes.

More surprisingly, we found that the second codon position is more conserved on average than the first one. A possible reason for this difference is that while three amino acids (Leucine, Serine and Arginine) are encoded by codons which differ on the first codon position, only one amino acid (Serine) is encoded by codons which differ on the second position. This can allow greater average flexibility in the first codon position, reflected in greater entropy.

Additionally, the amplitude of the periodicity seems to grow along the gene, as the third codon position has smaller variability in the first codons than further down the gene. Previous studies have suggested that

codon usage bias at the beginning of the gene can have an important influence on gene expression level and energetic cost of production (Frumkin et al., 2017) and that codon preferences are the strongest at the beginning of the gene (Kelsic et al., 2016), thereby resulting in lower entropy in this region.

As expected, in non-coding genes this periodicity pattern does not exist (Fig. 6, bottom). This observation can be used in future analyses of the genes which we described above to have very low variability, while their function is not yet known, and may suggest whether they are RNA genes or harbor the entropy patterns of protein coding genes.

Genomic entropy landscape across the human gut microbiome

Out next goal was to explore the genomic entropy landscape of different bacterial species in different populations of human gut microbiome to learn about different factors associated with bacterial genomic variability. For this, we have applied the framework we created on all microbiome samples in our cohort and on all bacterial species we were able to detect.

Since different bacteria share genomic regions, GEM assigned more than one mapping destinations for some reads. Correct read assignment is critical for analyzing intra-species variability, without introducing inter-species variability caused by incorrect mapping. To address this issue, we have further analyzed the output from GEM with another algorithm which assigns probabilities to each mapping destination (Methods) and included only reads with high probability for correct mapping. There is a trade-off between setting the probability threshold too low, which increases the inter-species 'noise' caused by mis-mapping, to setting the threshold too high which will result in often ignoring regions which are similar between two species and thus missing relevant information.

To choose a threshold for mapping destination probability, we examined the distribution of the most probable mapping destination reads from 10 of the samples. As an intermediate, we chose a probability threshold of 0.99 for inclusion of reads, which allowed us to include most reads (89.54% of 3.04×10^7 sampled reads) while still having confidence in their genome assignments.

We have mapped all samples' sequenced reads to known bacterial genomes and generated a nucleotide frequency array for each genome in each sample (Methods), creating a total of 237,025 arrays, one for each existing combination of microbiome sample × bacterial species. To accurately sample the entropy of each position, in each sample we only included positions which were covered by 5 reads or more. For initial analyses, in order increase the coverage of each genome (and thus increase the information of each position) and the number of samples covering each species, we merged all samples of each participant together. To be able to calculate the average entropy of a genome and compare different genomes with some confidence, we needed to choose a threshold for the minimal number of covered positions in a genome in order to deduce its average entropy. We then examined the distribution of the number of positions that meet the 5 reads per

position demand in each genome, and chose to include only genomes with at least 10,000 positions that meet this demand, as a balance between including only high-covered genomes to increase the accuracy of calculating the mean entropy, and including less-covered genomes to increase our power to observe general phenomena over a larger number of genomes (Fig. 7a).

After this filtering step, we were left with 44,574 genomes from 1003 human hosts and 686 bacteria species, with the number of different species in a sample ranging from 1 to 110, and the number of different samples per species ranging from 1 to 979 (Fig. 7b,c). In all following analyses, we have only included samples with 15 bacterial species or more, and species which are sampled 40 times or more, leaving us with 43,253 genomes to analyze.

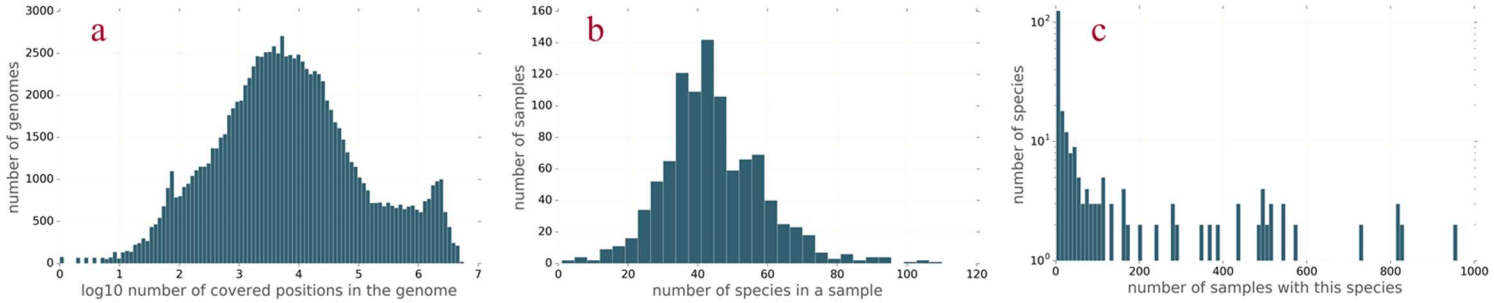

Figure 7 – a. number of positions which are covered by 5 reads or more, in all genomes. We chose a threshold of 10,000 covered positions or more. b. number of species detected in each sample. c. number of different samples in which each species was found.

As first step, we examined whether the observed genomic variability is associated with the bacteria itself, with some characteristics of the human host, or with both. Visualizing the average entropy in each genome (Fig. 8), it appeared that neither the species nor the host solely determine the genomic variability of the bacteria.

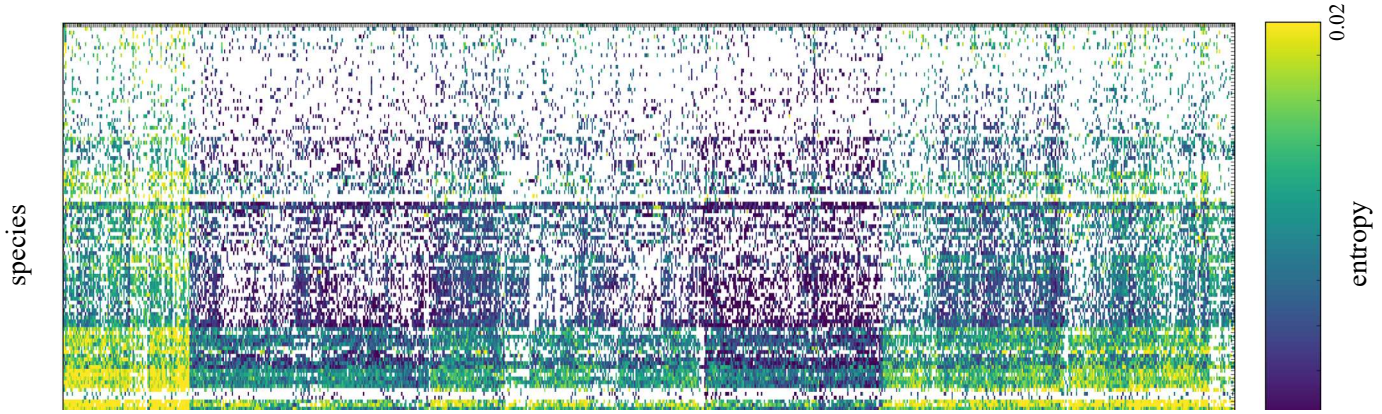

Figure 8 – **Mean genomic entropy varies between hosts and bacterial species** heatmap showing the mean entropy of each genome, which represents one bacterial species in one sample. The matrix is redundant (white) since not all species were present in each sample. To reduce the redundancy of the visualization, only species which exists in 60 samples or more, and only samples which cover 30 bacterial species or more, are included.

Plotting the mean entropy of each host and the mean entropy of each species, show a wide range in both distributions (Fig. 9).

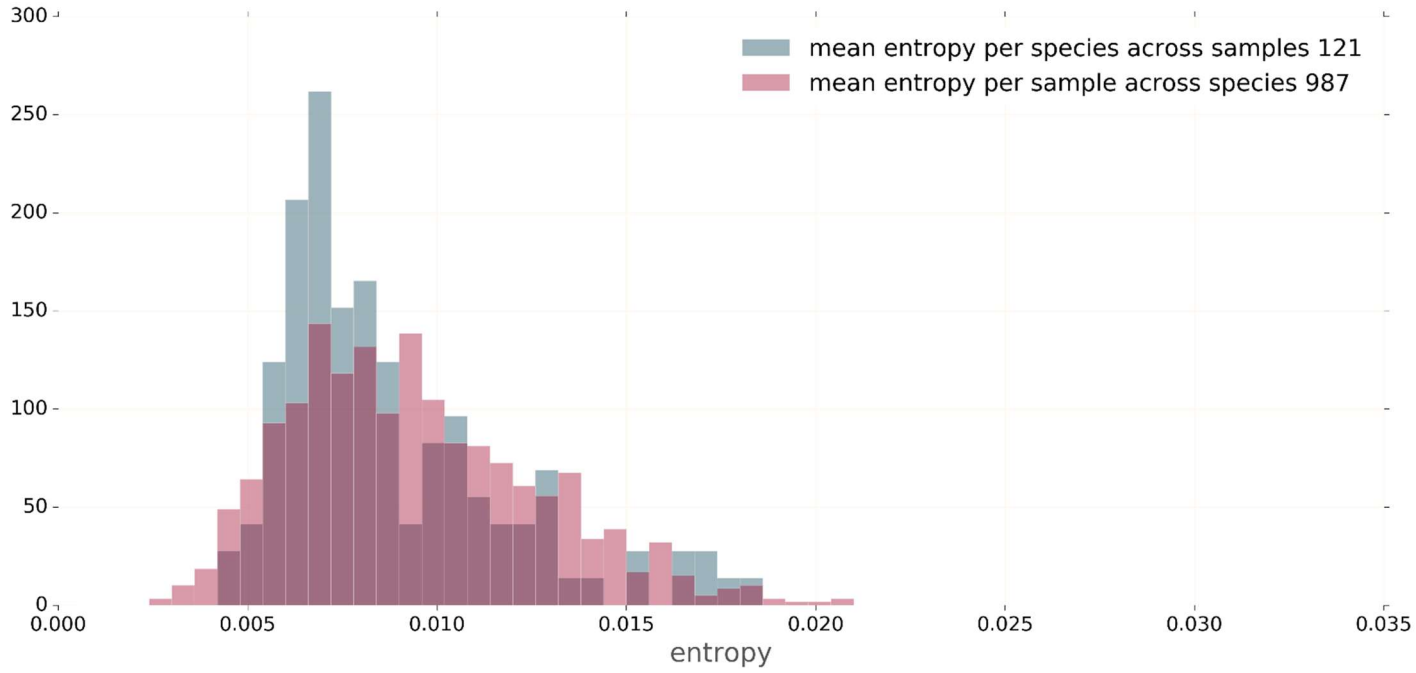

Figure 9- **Mean genomic entropy varies between hosts and bacterial species**

Bacterial factors associated with genomic variability

To isolate species factors associated with the genomic variability from factors of the sample or the host, after calculating the average entropy for each genome, we calculated each species' entropy score (Methods).

We next sought to examine the association of genomic variability and species phylogeny.

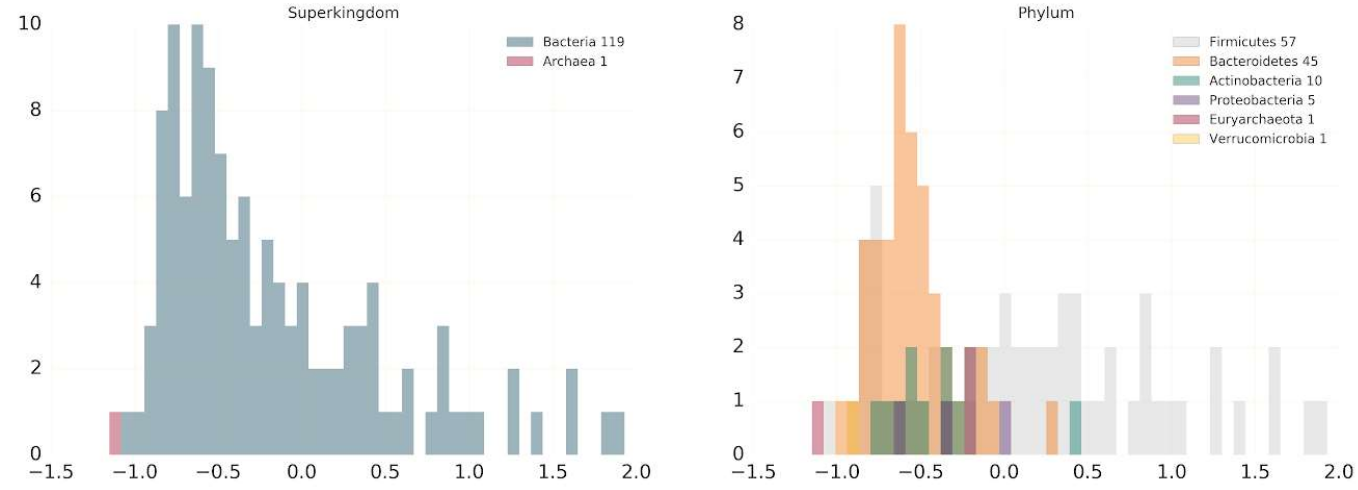

Figure 10 – **Genomic variability is taxa-dependent** entropy score of all species (left) Entropy scores distribution is different for different superkingdoms, Mann-Whitney-U test p = 0.044. (right) Entropy scores distribution is different for different phylum. Mann-Whitney-U test p for the separation between Firmicutes and Bacteroidetes < 10^{-6}

We found that the overall genomic entropy score is highly taxa-dependent (Fig. 10). We hypothesized that the such difference may result from different mutation rates, evolvement of different repair mechanisms or different selective pressures on different taxa groups. Future studies may call for the comparison of DNA polymerases and other components of DNA replication/repair processes, to see whether these can predict species-wise genomic variability. A possible artifact that may affect this result, is the quality of the reference genomes. If the precision in genome sequences is highly different for different taxa, there may be difference in the accuracy of read assignment, resulting in perceived genomic variability. To test that, a thorough examination of the reference dataset is needed, to see whether the number of sequenced genomes of each species in a taxon inversely correlates with the observed genomic variability.

As another bacterial factor, we tested whether the amount of variability in a bacterium's genome is correlated with its length, which might have suggested that larger genomes are more redundant and thus may be more robust to some mutations. However, we did not observe a significant correlation between genome size and variability score (Spearman's $P = 0.26$).

Host factors associated with genomic variability in the microbiome

We were very curious to understand what distinguishes people with high entropy score of their gut microbiome from people with low microbiome entropy scores, as these association might improve our understanding of host-microbiome interaction. We hypothesized that certain medical phenotypes and habits will be associated with increased or decreased genomic variability in the residing bacteria.

Similarly to what we did in the analysis of the species factors, we started by calculating the entropy score for each sample (Methods), to eliminate the effect of entropy differences between species. To minimize the likelihood of technical artifacts, we included a single sample per participant in this and in all following analyses (Methods) resulting in a total of 656 samples for the analysis, representing 656 cohort participants. Next, we compared the hosts' entropy scores with about 30 host phenotypes and lifestyle habits, hoping to obtain a correlation between host characteristics and overall microbiome entropy (Fig.11, 12).

We found several phenotypes to be correlated with the entropy score, and these were only marginally significant upon multiple hypotheses correction. Interestingly, we found no statistically significant correlation between the entropy score of a sample and its alpha diversity. The former measures the intra-species genomic diversity in a sample, while the latter measures the species composition diversity in a sample. This result implies that these two levels of variability are independent.

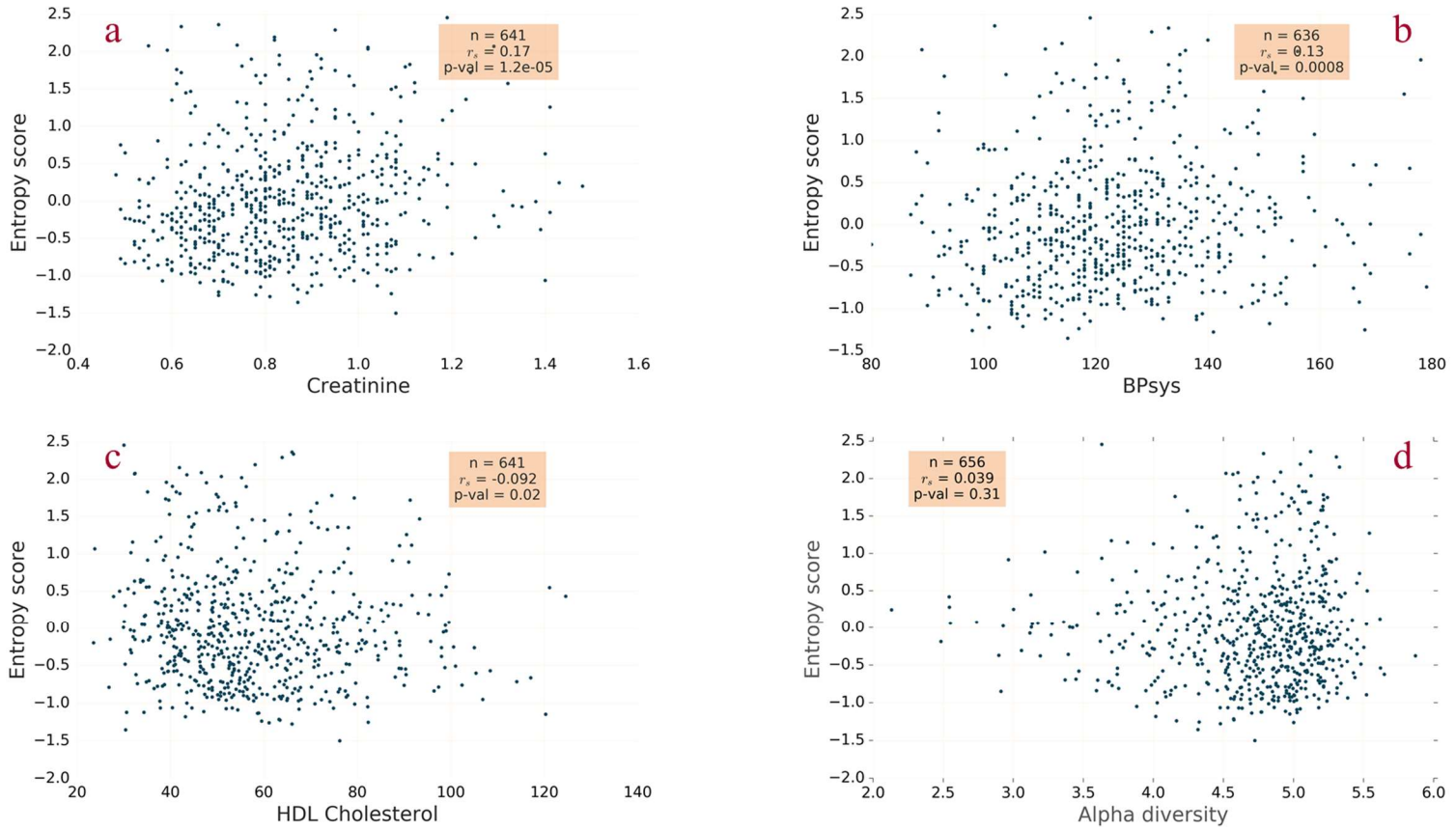

Figure 11 - **Several host phenotypes correlate with entropy score** in the squares: number of samples included in each analysis, Spearman's rank correlation coefficient and p-value. a. host creatinine levels vs. its entropy score b. host systolic blood pressure vs. its entropy score c. host HDL cholesterol vs. its entropy score d. host microbiome alpha diversity vs. its entropy score

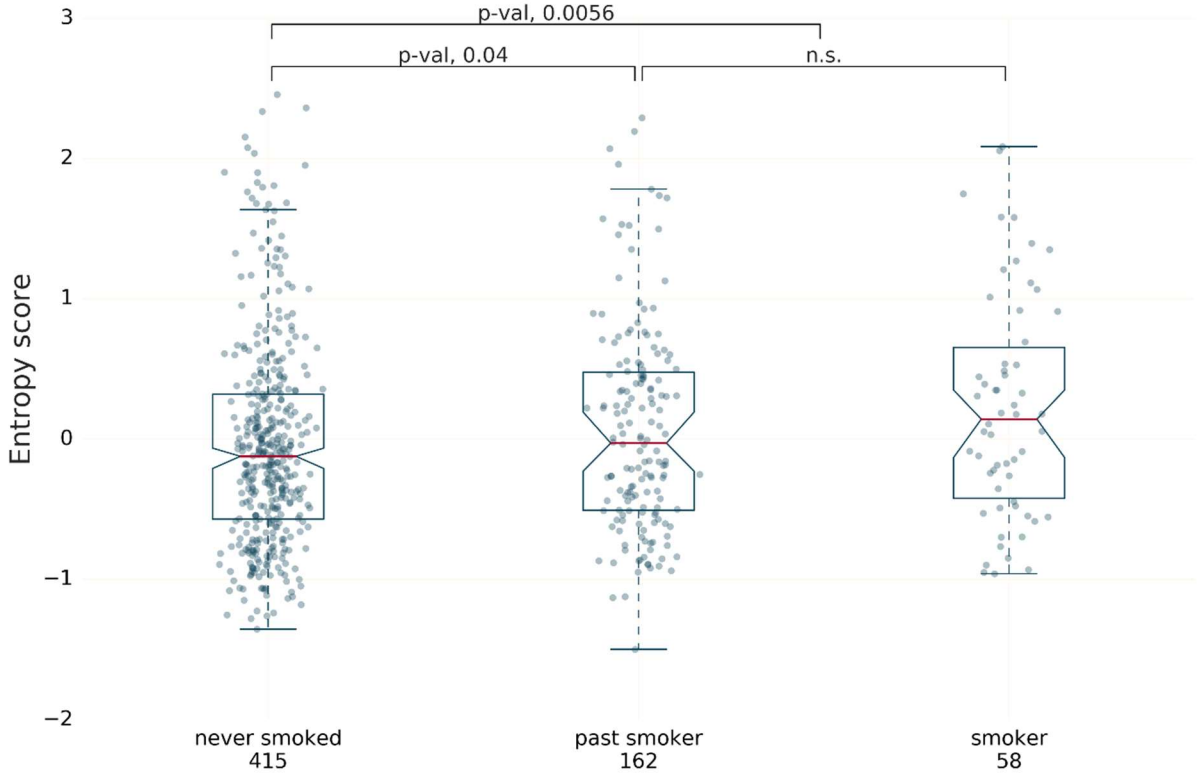

Figure 12 – **Smoking habit is associated with higher entropy score** comparison between entropy scores of people who never smoked, people who used to smoke and people who currently smoke. P-value for the null hypothesis that distributions are the same was calculated with Mann-Whitney-U-rank test. Boxplot notches mark the 95% confidence interval for the median, based on bootstrapping of 1000 repetitions

Next, we sought to examine the association of the genomic variability of different species and host features. We were highly curious to examine the association with smoking, therefore for each of the 93 bacterial species considered, we compared the average genomic entropy distribution in smokers' with the entropy distribution in samples of hosts who have never smoked, and tested whether these distributions vary (Fig. 13). After adjusting false discovery rate at 0.05, four species exhibited a significant association with smoking.

Interestingly, we observe that almost all species had consistently higher entropy values in smokers as compared to samples from individuals who never smoked. However, having a small effect in each species separately, and relatively smaller number of smokers in the cohort (one bacterial species was only observed in one smoker sample) made it difficult to observe a statistically significant difference within a species.

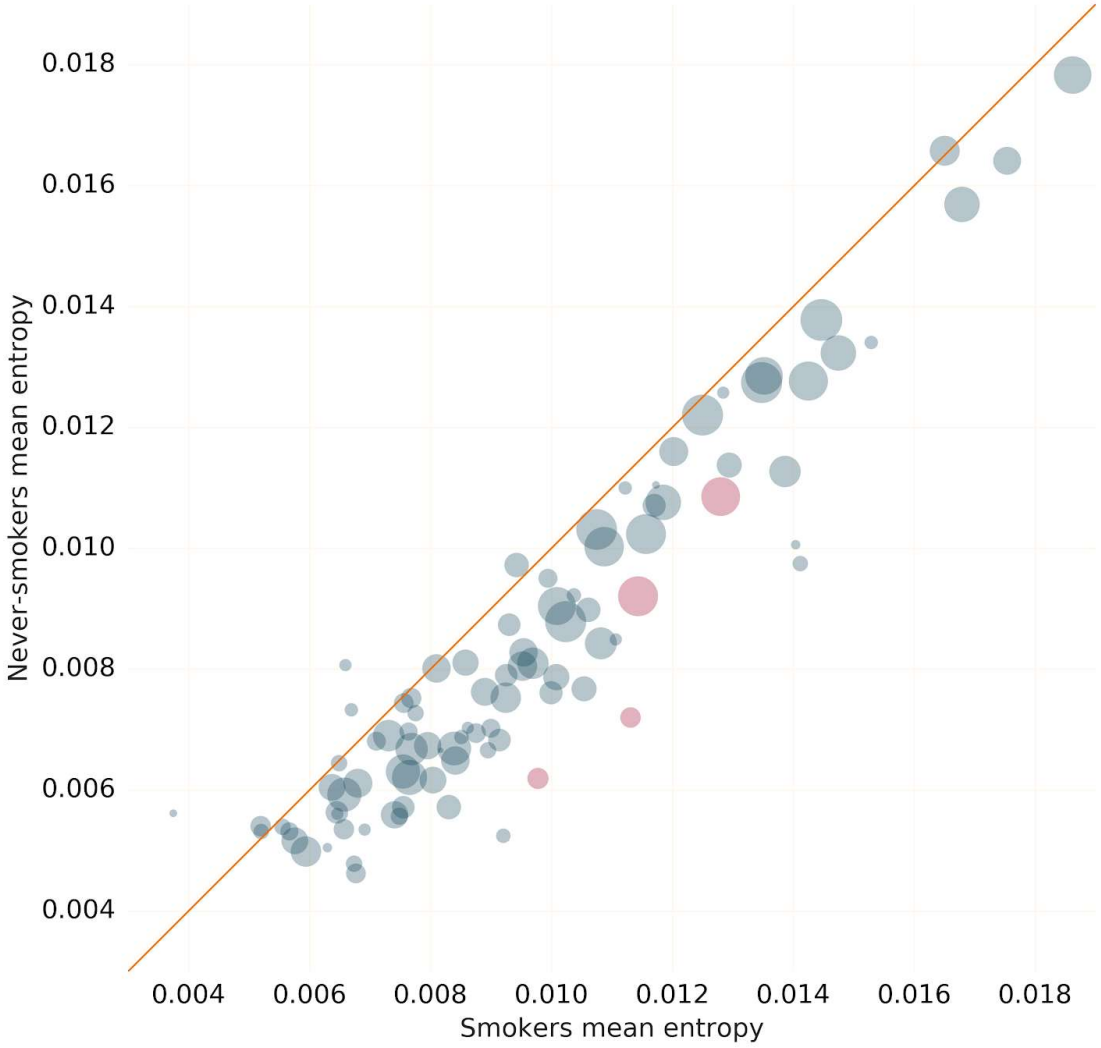

Figure 12 - **species-wise average entropy in smokers vs. its entropy in never-smokers** size of the bubble corresponds to the number of smokers-samples containing this species. Species with FDR-adjusted Mann Whitney $p < 0.05$ in pink.

We next wanted to test whether the overall correlation of genomic variability and smoking, when summed in each of the species, is bigger than what can be expected by chance. To test that, we used Wilcoxon signed-rank test over all species (each is a pair of smokers' average and non-smokers' average). We repeated this with additional binary phenotypes. To test the association of numeric host phenotypes and species entropy, we calculated for each species separately the correlation (Spearman) between its entropy in different hosts and the value of the phenotype in these hosts. Then, we counted how many species had a positive correlation (not necessarily statistically significant) with the phenotype, and compared this number with its probability in a binomial distribution (thinking of each species as a coin flip, with an equal probability of its correlation to be positive or negative, and counting how many of the 'coin flips' turned positive) to estimate whether the direction of the association is consistent across all species. For each binary phenotype we analyzed only species with at least 20 samples in each phenotype group ('yes' and 'no' groups) and for each numeric phenotype we analyzed species with at least 20 different hosts which had a valid value for this phenotype (Table 1).

Phenotype	Number of species with positive correlation	Total number of species considered	FDR adjusted p of the correlation between entropy and phenotype	Correlation sign
WBC	90	93	5.4E-22	+
Eosinophils %	8	93	2.3E-16	-
Tel-Aviv? (location)	91	91	7.9E-16	+
HDL Cholesterol	11	93	5.7E-14	-
Sodium	82	93	5.7E-14	+
Blood pressure, systolic	81	93	3.3E-13	+
Heart rate	80	93	1.8E-12	+
Ever smoked?	68	83	1.4E-11	+
Basophils %	76	93	9.9E-10	+
Calcium	18	93	3.8E-09	-
Currently smokes?	46	48	6.3E-09	+
Age	25	93	1.4E-05	-
Blood pressure, diastolic	68	93	1.4E-05	+
Hemoglobin	66	93	9.2E-05	+
ALT (Alanine aminotransferase)	31	93	2.3E-03	-
Neutrophils %	56	93	7.7E-02	+
Body weight	40	93	2.4E-01	-
HbA1C%	40	93	2.4E-01	-
Lymphocytes %	51	93	4.3E-01	+
Monocytes %	45	93	8.4E-01	-

Table 1 – **Correlation between genomic variability of the microbiome species and host phenotype** positive correlation sign for binary phenotypes means that the entropy of these species in the microbiome of hosts who 'have' this phenotype (tested in Tel-Aviv/ ever smoked/ currently smoke) is higher than in the microbiome of hosts who 'don't have' this phenotype.

Many of the host phenotypes have a statistically significant correlation with per-species entropy. White Blood Cells count (WBC) is one statistically significant example: in 90 out of 93 species we found some positive correlation between the species entropy and WBC. WBC is a marker for immune activity, was established as a risk factor for coronary heart disease and type 2 diabetes and is associated with smoking and obesity (Friedman, Siegelaub, Seltzer, Feldman, & Collen, 1973; Hoffman, Blum, Baruch, Kaplan, & Benjamin, 2004; Nieman et al., 1999; Vozarova et al., 2002). An association between high immune activity and higher bacterial genomic variability, can suggest interaction between the host immune system and the bacterial genome.

A surprising result is the correlation between bacterial entropy and the meeting location of the participants in the cohort. Data collection for the original study (Zeevi et al., 2015) was done in two different locations, Tel-Aviv or Rehovot, and each participant was assigned to one of the locations. A possible biological cause for this observation is different meeting hours in the two locations, which might correlate with different stool collection hours. Thaiss *et al.* (Thaiss et al., 2014) previously described diurnal oscillations in the composition and function of the gut microbiome, implying that different amount of genetic variability may be observed in different parts of the day. To test this hypothesis, it would be interesting to calculate the correlation between the time of day of stool collection and the measured genetic variability of bacteria. A more technical possible cause for this result can be a difference in sequencing error rate between different sequencing batches, which together with a non-uniform assignment of samples of both centers to sequencing batches might result in an apparent association between calculated entropy and meeting location. These two examinations are needed to reassure that no hidden confounders are affecting our results.

As we observed a correlation between smoking and bacterial entropy, we wanted to shed light on the source of this correlation. We present two competing hypotheses for higher entropy in smokers' microbiome species: (1) smoking introduces new selective pressure in the gut, which encourages positive selection; or (2) smoking induces random mutagenesis, similar to that observed in the effect of smoking on the host (Pleasance et al., 2010). These hypotheses suggest that smoking affects the genomic variability, and not vice versa, although this effect cannot be rolled-out based on existing data, as well. To test the two hypotheses, we quantified the genetic variability of these species in each codon position separately.

Due to the nature of the genetic code, the first two positions of the codon are expected to be less variable than the third, as substitutions in these positions are more likely to be nonsynonymous than a substitution in the third position. We have seen evidence for that in the periodicity pattern of entropy in *A. muciniphila* protein coding genes. Thus, we used the ratio between the entropy of the first two codon positions and the entropy of the third codon position, as an approximation for the established pN/pS ratio (McDonald & Kreitman, 1991; Schloissnig et al., 2013; Simmons et al., 2008) for nonsynonymous to synonymous mutations. The ratio was calculated for each genome (each species and each sample) separately (Methods).

We observed that the ratio between the entropy of the first two codon positions and the third is greater on average in smokers than in people who never smoked (Wilcoxon signed-rank test p-value = 0.0013, Fig. 14). When comparing the ratio between smokers mean entropy and never-smokers mean entropy in each of the codon positions separately, it shows that while all positions have higher average entropy in smokers' microbiome, the difference between groups is biggest in the second position, and significantly smaller in the third codon position (Fig. 15). So not only do all codon positions have higher entropy in smokers' microbiome, the ratio between the first two and the third codon positions is bigger in smokers as well. The combination of both results suggests positive selection, as more variation is likely nonsynonymous and might have adaptive value to an environmental challenge.

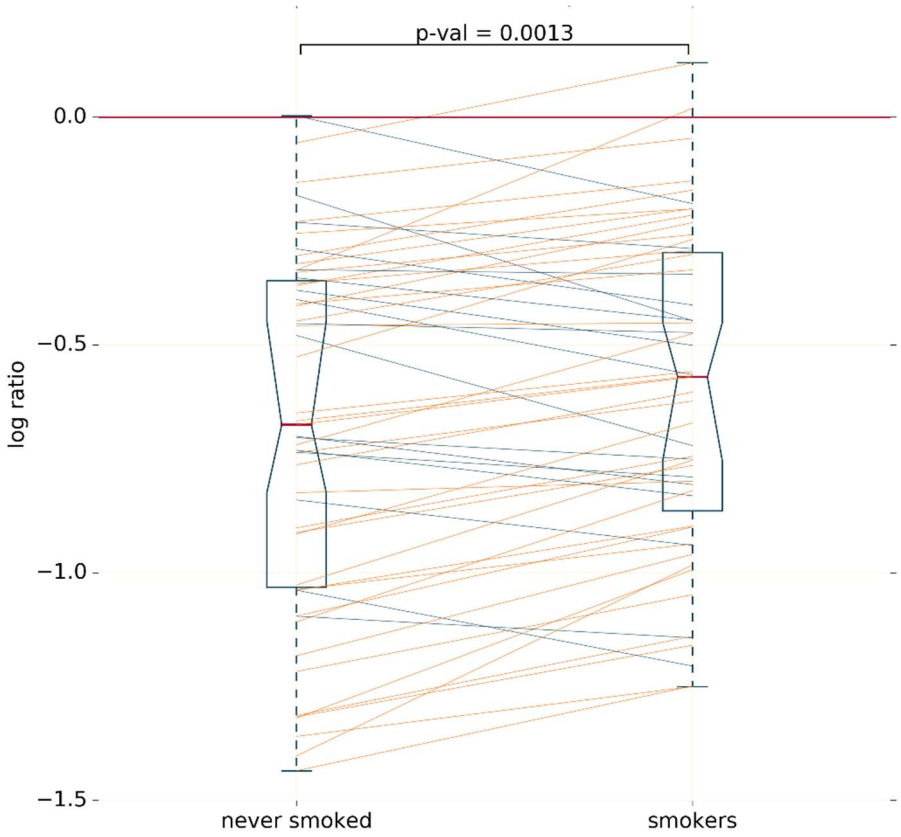

Figure 13 – **The ratio between the first two codon positions and the third is bigger in smokers microbiome** Y-axis is log2(ratio). Each line corresponds to one species, connecting its average ratio for participants who have never smoked with its average ratio in smokers. Blue species are those who have higher average ratio for never-smokers. Boxplot notches mark the 95% confidence interval for the median, based on bootstrapping of 1000 repetitions.

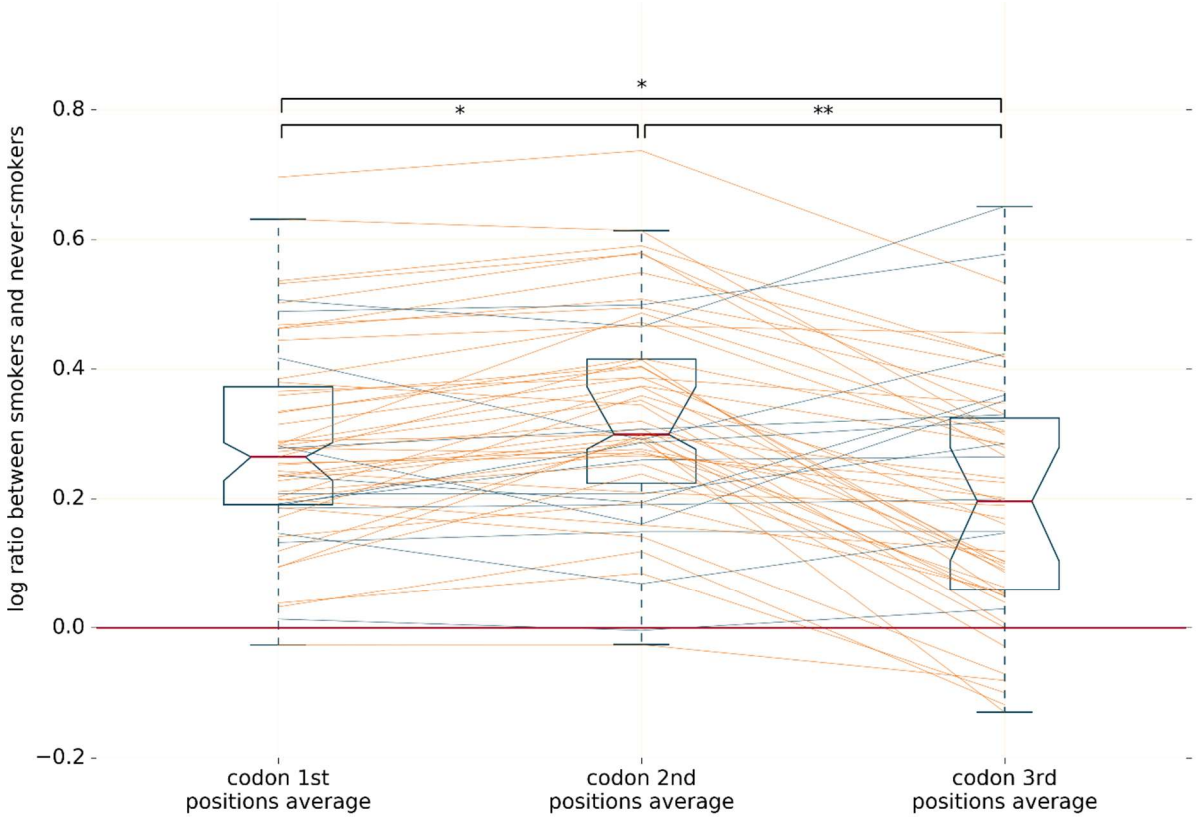

Figure 14 – **The observed association of smoking and genomic variability is largest in the second codon position** Y-axis is the log2 of the ratio between the species' mean entropy in smokers and its mean entropy in never-smokers, for each codon position separately. Each line corresponds to one species, connecting its average entropy for different hosts groups. Blue species are those with the biggest ratio at the third codon position. * Wilcoxon signed rank test $p < 0.00005$, ** Wilcoxon signed rank test $p < 0.000005$. Boxplot notches mark the 95% confidence interval for the median, based on bootstrapping of 1000 repetitions.

Discussion

In this project, we have devised a computational framework to analyze genomic variability in the human gut microbiome based on metagenomic samples, on a single nucleotide resolution. We described the genomic variability landscape of the gut bacterium *A. muciniphila* across a cohort of 1003 participants and compared the genetic diversity of different regions within its genome. We next analyzed the overall variability in the genomes of 93 bacterial species in microbiome samples of 656 hosts, observing a range of variability values which changes between hosts and species, often in association with host and bacterial features.

We have compared the entropy of different regions in the genome of *A. muciniphila*, and found that genes with different functions have different variability patterns. We observed a significantly lower entropy in RNA genes, a periodicity pattern in the entropy along protein coding genes, and a distinguished pattern of a CRISPR array. As many of the bacterial genes are still labeled as 'hypothetical proteins', in the future, the intra-species variability patterns may be used to predict the function of genomic regions and discover open reading frames in unannotated genomes.

We found that some of the most variable genes in the genome of *A. muciniphila* have domains of membrane and transporter proteins. This corresponds with a study of intra-host evolution of the commensal gut species *Bacteroides fragilis,* which revealed that most of its genes under positive selective pressure in the human gut are either outer membrane importers, or proteins involved in cell envelope synthesis (Zhao et al., 2017). These genes can potentially be key in adaptation to host immune system, phage diversity and nutrients availability, highlighting them for future analysis of bacterial adaptation to the host gut. We have also found that high genomic variability in the gut microbiome is positively associated with host white blood cells counts, a marker for the activity of the immune system. These two observations suggest that host immune activity may create a challenge to the symbiont bacteria which promotes variability toward adaptation. Further analyzing whether the association between bacterial variability and host immune activity results from high variability of specific genomic regions may reveal bacterial genes which directly or indirectly interact with the immune system. Furthermore, while in this project we have demonstrated extreme variability in these genes across the entire cohort, it would be interesting to test whether the source to this overall variability is between-host variance, while the within-host variability of these genes remains small. This may imply that different variants of these genes fit bacteria in different hosts' guts, and examination of these variants may promote our understanding of the interaction between the microbiota and the human immune system.

Integrating both parts of the project: comparing the variability in different parts of the genome and analyzing the bacterial variability in the context of host phenotypes and lifestyle habits, may highlight genomic regions which are important for bacterial survival and adaptation in specific conditions, and promote our understanding of bacterial adaptation mechanisms and the interaction between the host and the microbiome.

The diversity of species in the microbiome has been found as beneficial to the host (Turnbaugh et al., 2009), however, genetic diversity of the microbiome exists in additional resolutions. Genetic diversity can be achieved by variability of species abundance, as well as by intra-species variation of gene copy numbers, mobile genes composition and single nucleotide polymorphisms- all have been previously demonstrated to exist in the gut microbiome (Brito et al., 2017; Greenblum, Carr, & Borenstein, 2015; Schloissnig et al., 2013; Turnbaugh et al., 2009). We found that the diversity of species and the amount of intra-species variability in the microbiome do not correlate, while observing correlation between the latter and a few host phenotypes. We therefore suggest using the overall amount of intra-species variability as another feature that characterizes one's microbiome and call for future examination of its association with host health.

Literature

Agarwala, R., Barrett, T., Beck, J., Benson, D. A., Bollin, C., Bolton, E., … Zbicz, K. (2017). Database Resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, *45*(D1), D12–D17. https://doi.org/10.1093/nar/gkw1071

Altshuler, D. L., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., … Peterson, J. L. (2010). A map of human genome variation from population-scale sequencing. *Nature*, *467*(7319), 1061–1073. https://doi.org/10.1038/nature09534

Brito, I. L., Yilmaz, S., Huang, K., Xu, L., Jupiter, S. D., Jenkins, A. P., … Alm, E. J. (2017). Mobile genes in the human microbiome are structured from global to individual scales. *Nature*, *544*(7648), 124–124. https://doi.org/10.1038/nature20774

Clayton, T. A., Baker, D., Lindon, J. C., Everett, J. R., & Nicholson, J. K. (2009). Pharmacometabonomic identification of a significant host-microbiome metabolic interaction affecting human drug metabolism. *Proceedings of the National Academy of Sciences*, *106*(34), 14728–14733. https://doi.org/10.1073/pnas.0904489106

Crick, F. H. C. (1966). Codon-anticodon pairing : the wobble hypothesis. *Journal of Molecular Biology*, *19.2*, 548–555. https://doi.org/https://doi.org/10.1016/S0022-2836(66)80022-0

David, L. A., Materna, A. C., Friedman, J., Campos-Baptista, M. I., Blackburn, M. C., Perrotta, A., … Alm, E. J. (2015). Host lifestyle affects human microbiota on daily timescales. *Genome Biology*, *15*(7), 1–15. https://doi.org/10.1186/gb-2014-15-7-r89

Frank, D. N., St Amand, A. L., Feldman, R. A., Boedeker, C. E., Harpaz, N., & Pace, N. R. (2007). Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. *Proc Natl Acad Sci USA*, *104*. https://doi.org/10.1073/pnas.0706625104

Friedman, G. D., Siegelaub, A. B., Seltzer, C. C., Feldman, R., & Collen, M. F. (1973). Smoking Habits and the Leukocyte Count. *Archives of Environmental Health: An International Journal*, *26*(3), 137–143. https://doi.org/10.1080/00039896.1973.10666241

Frumkin, I., Schirman, D., Rotman, A., Li, F., Zahavi, L., Mordret, E., … Pilpel, Y. (2017). Gene Architectures that Minimize Cost of Gene Expression. *Molecular Cell*, *65*(1), 142–153. https://doi.org/10.1016/j.molcel.2016.11.007

García-Arenal, F., Fraile, A., & Malpica, J. M. (1999). Genetic Variability and Evolution. In *Molecular Biology of Plant Viruses* (pp. 143–159). Boston, MA: Springer US. https://doi.org/10.1007/978-1-4615-5063-1_6

Greenblum, S., Carr, R., & Borenstein, E. (2015). Extensive Strain-Level Copy-Number Variation across Human Gut Microbiome Species. *Cell*, *160*(4), 583–594. https://doi.org/10.1016/j.cell.2014.12.038

Grissa, I., Vergnaud, G., & Pourcel, C. (2007). CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Research*, *35*(Web Server), W52–W57. https://doi.org/10.1093/nar/gkm360

Hoffman, M., Blum, A., Baruch, R., Kaplan, E., & Benjamin, M. (2004). Leukocytes and coronary heart disease. *Atherosclerosis*, *172*(1), 1–6. https://doi.org/10.1016/S0021-9150(03)00164-3

Hong, C., Manimaran, S., Shen, Y., Perez-Rogers, J. F., Byrd, A. L., Castro-Nallar, E., … Johnson, W. E. (2014). PathoScope 2.0: A complete computational framework for strain identification in environmental or clinical sequencing samples. *Microbiome*, *2*(1), 1–15. https://doi.org/10.1186/2049-2618-2-33

Hsiao, E. Y., McBride, S. W., Hsien, S., Sharon, G., Hyde, E. R., McCue, T., … Mazmanian, S. K. (2013). Microbiota modulate behavioral and physiological abnormalities associated with neurodevelopmental disorders. *Cell*, *155*(7), 1451–1463. https://doi.org/10.1016/j.cell.2013.11.024

Kelsic, E. D., Chung, H., Cohen, N., Park, J., Wang, H. H., & Kishony, R. (2016). RNA Structural Determinants of Optimal Codons Revealed by MAGE-Seq. *Cell Systems*, *3*(6), 563–571.e6. https://doi.org/10.1016/j.cels.2016.11.004

Korem, T., Zeevi, D., Zmora, N., Weissbrod, O., Bar, N., Lotan-Pompan, M., … Segal, E. (2017). Bread Affects Clinical Parameters and Induces Gut Microbiome-Associated Personal Glycemic Responses. *Cell Metabolism*, *25*(6), 1243–1253.e5. https://doi.org/10.1016/j.cmet.2017.05.002

Larsen, N., Vogensen, F. K., Van Den Berg, F. W. J., Nielsen, D. S., Andreasen, A. S., Pedersen, B. K., … Jakobsen, M. (2010). Gut microbiota in human adults with type 2 diabetes differs from non-diabetic adults. *PLoS ONE*, *5*(2). https://doi.org/10.1371/journal.pone.0009085

Lieberman, T. D. (2018). Seven Billion Microcosms: Evolution within Human Microbiomes. *mSystems*, *3*(2), e00171-17. https://doi.org/10.1128/mSystems.00171-17

Marco-Sola, S., Sammeth, M., Guigó, R., & Ribeca, P. (2012). The GEM mapper: Fast, accurate and versatile alignment by filtration. *Nature Methods*, *9*(12), 1185–1188. https://doi.org/10.1038/nmeth.2221

Martincorena, I., Roshan, A., Gerstung, M., Ellis, P., Van Loo, P., McLaren, S., … Campbell, P. J. (2015). High burden and pervasive positive selection of somatic mutations in normal human skin (Supplement). *Science*, *348*(6237), 880–886. https://doi.org/10.1126/science.aaa6806

McDonald, J. H., & Kreitman, M. (1991). Adaptive protein evolution at the Adh locus in Drosophila. *Nature*, *351*(6328), 652–654. https://doi.org/10.1038/351652a0

Mende, D. R., Letunic, I., Huerta-Cepas, J., Li, S. S., Forslund, K., Sunagawa, S., & Bork, P. (2017). ProGenomes: A resource for consistent functional and taxonomic annotations of prokaryotic genomes. *Nucleic Acids Research*, *45*(D1), D529–D534. https://doi.org/10.1093/nar/gkw989

Nieman, D. C., Henson, D. A., Nehlsen-Cannrella, S. L., Ekkens, M., Utter, A. C., Butterworth, D. E., & Fagoaga, O. R. (1999). Influence of Obesity on Immune Function. *Journal of the American Dietetic Association*, *99*(3), 294–299. https://doi.org/10.1016/S0002-8223(99)00077-2

Pleasance, E. D., Stephens, P. J., O'Meara, S., McBride, D. J., Meynert, A., Jones, D., … Campbell, P. J. (2010). A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature*, *463*(7278), 184–190. https://doi.org/10.1038/nature08629

Schloissnig, S., Arumugam, M., Sunagawa, S., Mitreva, M., Tap, J., Zhu, A., … Bork, P. (2013). Genomic variation landscape of the human gut microbiome. *Nature*, *493*(7430), 45–50. https://doi.org/10.1038/nature11711

Sender, R., Fuchs, S., & Milo, R. (2016). Revised Estimates for the Number of Human and Bacteria Cells in the Body. *PLoS Biology*, *14*(8), 1–14. https://doi.org/10.1371/journal.pbio.1002533

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(July 1928), 379–423. https://doi.org/10.1145/584091.584093

Simmons, S. L., DiBartolo, G., Denef, V. J., Aliaga Goltsman, D. S., Thelen, M. P., & Banfield, J. F. (2008). Population genomic analysis of strain variation in Leptospirillum group II bacteria involved in acid mine drainage formation. *PLoS Biology*, *6*(7), 1427–1442. https://doi.org/10.1371/journal.pbio.0060177

Sorek, R., Kunin, V., & Hugenholtz, P. (2008). CRISPR - A widespread system that provides acquired resistance against phages in bacteria and archaea. *Nature Reviews Microbiology*, *6*(3), 181–186. https://doi.org/10.1038/nrmicro1793

Thaiss, C. A., Zeevi, D., Levy, M., Zilberman-Schapira, G., Suez, J., Tengeler, A. C., … Elinav, E. (2014). Transkingdom control of microbiota diurnal oscillations promotes metabolic homeostasis. *Cell*, *159*(3), 514–529. https://doi.org/10.1016/j.cell.2014.09.048

Turnbaugh, P. J., Hamady, M., Yatsunenko, T., Cantarel, B. L., Duncan, A., Ley, R. E., … Gordon, J. I. (2009). A core gut microbiome in obese and lean twins. *Nature*, *457*(7228), 480–484. https://doi.org/10.1038/nature07540

Tyson, G. W., & Banfield, J. F. (2008). Rapidly evolving CRISPRs implicated in acquired resistance of microorganisms to viruses. *Environmental Microbiology*, *10*(1), 200–207. https://doi.org/10.1111/j.1462-2920.2007.01444.x

Vozarova, B., Weyer, C., Lindsay, R. S., Pratley, R. E., Bogardus, C., & Tataranni, P. A. (2002). High white blood cell count is associated with a worsening of insulin sensitivity and predicts the development of type 2 diabetes. *Diabetes*,

51(2), 455–461. https://doi.org/10.2337/diabetes.51.2.455

Zeevi, D., Korem, T., Zmora, N., Israeli, D., Rothschild, D., Weinberger, A., … Segal, E. (2015). Personalized Nutrition by Prediction of Glycemic Responses. *Cell*, *163*(5), 1079–1095. https://doi.org/10.1016/j.cell.2015.11.001

Zhao, S., Lieberman, T. D., Poyet, M., Groussin, M., Gibbons, S. M., Xavier, R. J., & Alm, E. J. (2017). Adaptive evolution within the gut microbiome of individual people. *bioRxiv*, 208009. https://doi.org/10.1101/208009

Acknowledgments

I have learned and grown a lot in the past couple of years, thanks to the brilliant, great people who surrounded me. I would like to acknowledge the contribution of the following people to my work and thank them:

To Eran, for believing in me and giving me this opportunity to join his research group of exceptionally talented people. I appreciate how he allows me immense independence while also always being available for help or discussion. The example he sets for doing excellent and productive research has motivated me to challenge myself and strive to improve through every step of the way, and will definitely inspire me in future research as well. I learn so much about so many different topics and gain so many new skills from our discussions, from our shared work, and from many different forums that he has created in the lab.

To Tzachi, for the most interesting, stimulating and inspiring discussions. I love discussing evolution, biology and science in general with him, and I always leave our meetings with many exciting ideas, renewed motivation and strengthened passion for science. These have left a mark on who I am and who I aspire to be as a scientist. I also thank him for accompanying me in the process of writing this thesis, from which I have learned more than I have imagined.

To Dudi, for meaningful mentorship, for conceiving this project, for walking me through the first steps, and for advising me in many of the following ones, including in writing this thesis. His share in everything that I have learned working on this project is significant.

To Tal, for *always* being there for me when I need help, even though there are always 10^6 people who need it as well, and always being very helpful when you do.

Tal and Dudi have both become my academic big brothers, being role models in so many ways, providing valuable advice, constructive criticism and arguably funny jokes. Also, their work and their code have been the most important foundation for this project. I am grateful for that, and for the good friends they are.

I thank Iris, Adi and Noam, my desk-mates, and Hagai, who could have been. I was fortunate to share a desk with the kindest and smartest friends, from which I learn a lot every day. They are great friends and make everything much more enjoyable. I thank Adi for great helpful discussions regarding this project. I thank Hagai for reminding me to always smile and say ‘Good Morning’, even if I am too busy. I thank Iris for being the best M.Sc. partner I could have asked for.

To all members of the Segal Lab and of the Pilpel Lab, I thank for fruitful GM discussions, countless coffee breaks, friendly conversations and teachable moments. And for the much-needed supply of chocolate.

I am most grateful for the endless support from Ben. I thank him for helping me work, and not work, in so many amazing ways, and for helping me keep my sanity when in danger.

To my family, and most of all to my parents, for the great inspiration and continuous encouragement that have undoubtedly shaped the person I am, and allow me to follow my aspirations, including the pursuit of this thesis. I am grateful for their infinite love and support, they are the best in the world.