Inferring positive selection and demography in the human gut microbiota.

Christos Zioutis

Master thesis for the graduate program: “Molecular Biology and Biomedicine”

Department of Biology

University of Crete

2015

Abstract

The animal gut is considered to be one of the most densely populated by microorganisms ecosystem. All this microbial weight, vital for host's health and homeostasis, has been co-evolving along with the gut structure, for millions years now. Recent studies on human gut microbiome have already demonstrated the genetic repertoire of this community and highlighted its genetic variation among individuals, not only in health states but also in various disease conditions. However, the evolutionary processes that drive the genetic diversity of the gut microbiome have not been thoroughly investigated. Here, we use metagenomic datasets from stool samples of 39 healthy US individuals to construct the genomic variation landscape of the most abundant bacterial species in the human gut. Our main goals are, the detection of events of strong positive selection, based on the site frequency spectrum (SFS) and the inference of a demographic model for each one of the bacterial species comprising the metagenomic sample. To infer the demographic scenarios we use an approximate computation bayesian technique based on coalescent simulations, taking into account the SFS. We find that the abundance or even presence of previously reported as highly abundant gut species varies among the individuals in our cohort. We report the most possible demographic scenarios for each bacterial strains, which, for most of the bacteria, comprises a recent, sharp population size decline. Furthermore, we report our findings on the detection of strong positive selection. Finally,, we report a plethora of difficulties in the estimation of SFS from metagenomic samples, such as the method for estimating allele frequencies from datasets with missing observations.

Introduction

Billions of years now microorganisms are dominating the Earth. It seems that the size does not matter as microbes seem to be the pillars that support the edifice of life. From regulating the global carbon cycle to the tremendous impact of their symbiosis on plant and animal health, they are the key players of the biosphere. The total ecological diversity of microbes that occupy a certain ecosystem is termed the microbiome. Our anthropocentric perspective of microbial world for many decades made us believe that microbes are enemies. That is a half-truth, because in reality, they are our best ally in the battlefield, despite that in many cases they can prove to be pathogenic. Indeed, we are super-organisms comprising of a small cohort of human cells and a quite larger cohort of symbiotic microbial cells. The microbial cells in our body (10^{13}–$10^{14)}$ is estimated to outnumber the human cells by a factor of ten and contain at least 100 times as many genes (10^{5}) as our own genome[1]. The diversity and abundance of microbes is distinct for the different body habitats and each habitat is characterized by one or a few signature taxa making up the plurality of the community and within-individual variation over time is consistently lower than between individuals variation. The uniqueness of each individual's microbial community thus appears to be stable over time[2].

The human gut microbiota

The gut microbiota represents the most densely populated and diverse ecosystem on Earth and reaches its highest cell concentration in the colon, with a density of 10^{12} CFU/g of intestinal content[3]. Initial postnatal microbial exposure occurs during and shortly after the birth, with the mother's vaginal and fecal microbiomes serving as the most important sources[4]. The phylogenetic diversity of the infant's gut microbiome increases gradually over time. It also demonstrates discrete steps of bacterial succession driven by life events such as nutritional stages (e.g weaning, table food) and antibiotic treatments. However, upon the third year, the gut microbiota becomes relatively stable in terms of taxa composition, a phenomenon that persists during the adult life[5] . It is known that over 1,000 bacterial species colonize the human gut, with the vast majority belonging to the phyla of Firmicutes and Bacteroidetes (90%), while Actinobacteria Proteobacteria, Fusobacteria, Cyanobacteria and Verrucomicrobia being less well represented[6]. The stability in composition reported by research studies, heavily depends on the phylogenetic resolution and the density of time series experiments and despite a core microbiome that is stable over time, it seems that a portion of it fluctuates throughout life[7]. Interestingly, when samples from two different colons, were compared to samples from different sampling sites on one individual, the differences were greater[8]. A recent metagenomic study of the genomic variation in the gut microbiome, demonstrated that although there are changes in species abundances, genomic variability similarity tends to be high for the same individual over the time period of a year and might serve as a fingerprint of an individual. The same study, using metagenomic samples from stool of 207 US individuals, identified 10,3 millions SNPs in 101 gut species and reported that i) up to 99% of the metagenomic reads was mapped to 66 gut bacterial species. The same study also proposed that ii) host conditions (such as diet, genetic differences, and immune tolerance) have a minor influence on the evolution of species compared to constraints common to the human population (such as gut physiology, anaerobic conditions and pH) and that iii) the source of variation in human gut microbial populations is less likely to be new mutations withing the host than the variation in the initial colonizing populations or transmissions from the environment[9].

Alumni for symbiosis

The importance of symbiosis with microbes can be easily demonstrated by four fundamental processes, crucial for human health, that are maintained and regulated by this interaction; (i) the metabolism of complex polysaccharides[10], (ii) the production of vitamins and amino-acids[6], (iii) the training of the immune system[11] and (iv) brain development and functions[12-14].

The human genome is capable of fully degrading a small subset of glycans – namely starch, lactose and sucrose. However, the dietary glycans that enter the gut are exceptionally diverse. Gut bacteria, however, have accumulated a huge variety of carbohydrate active enzymes (CAZymes) in their genomes that enable the efficient digestion of almost all plant and animal derived polysaccharides. The final metabolites of this fermentation process are mainly short-chain fatty acids (SCFAs – acetate, propionate, butyrate). Acetate and propionate are absorbed into the bloodstream and travel to the liver, where they are incorporated into lipid and glucose metabolism respectively, whereas butyrate represents the main energetic substrate for the colonic epithelial cells. These acids have been shown to play a pivotal role in host nutrition and energy homeostasis, controlling energy production, and storage as well as the appetite. Perturbations in the production of these metabolites have been associated with a couple of diseases, such as colorectal cancer and obesity[15,16]. Except for the diet derived polysaccharides, gut bacteria have evolved to metabolize very complex and diverse endogenous host glycans secreted by mucus[10]. They can also digest a mixture of complex human milk oligosaccharides (HMO) that are abundant in the breast milk of humans but not of other mammals. Most HMOs cannot be digested by human enzymes, suggesting that they guide the development of the infant gut microbiota, selectively feeding specific species[17-19].

There is an continuous interplay between the commensal gut microbes and our immune system. The gut microbiota are required for the normal generation and/or maturation of GALTs, immune structures in which antigens can be taken up and presented by antigen-presenting cells, serving as the key for tolerance or inflammation. IgA has been also shown to affect and be affected by the composition of gut microbiota[10]. IgA is an immunoglobulin produced in the mucosal tissues, including the intestine, and coats commensal bacteria or other soluble antigens, inhibiting their binding to the host epithelium[20]. There is also evidence that gut microbiota promote the development of T_H17 cells, a type of T cells that preferentially accumulates in the intestine and have a role in the development of autoimmune diseases[21]. Additionally, many studies have reported that commensal bacteria can inhibit pathogen colonization. There are many proposed mechanisms by which they accomplish it. The successful competition with other pathogens for the limited supply of nutrients in the intestine, the promotion of mucus production and the stimulation of production of antimicrobial substances from epithelial cells are thought to be the main mechanisms[10].

A prominent role for the gut microbiota has been proposed for the gut-brain interactions, too. Based on studies using germ-free animals, the gut microbiota appears to influence the development of emotional behavior, stress and pain modulation systems, and brain neurotransmitter systems. This is also shown by experiments of antibiotic and probiotic treatment of animals[22-23]. Current evidence suggests that multiple mechanisms, including endocrine and neurocrine pathways, may be involved in gut microbiota-to-brain signaling[14]. However, it is long been known that the brain can in turn alter microbial composition and behavior via the autonomous nervous system[24]. Microbes may benefit from communicating with the brain. Maybe they need us to be social, so that they can spread through the human population.

The mutualistic relationship between gut microbes and humans is based on the fact that we have access to a variety of food sources that we cannot digest and directly utilize as energy source. This variability in food sources is a consequence both of the mobility that characterizes human lifestyle, and the fluctuations in food availability due to insufficient supplies or seasonal restrictions. From the other hand, microbes that hijack the human intestinal bus, advantage from this diversity in our diet and thrive in the gut with the constraint to provide us with nutrients. Thus, the human microbiome must have been evolved to retain a degree of plasticity so that it can easily respond to perturbations in food sources. It has been shown recently, that plant-based or animal-based diets promote changes the gut microbiota composition and these adaptations to diet can be detected in 1-3 days[25,26]. A molecular study of diet-dependent microbiota dynamics in 14 overweight men revealed that the individual microbiota adapts its phylogenetic profile in response to the main types of fermentable carbohydrates. Interestingly, this study revealed subject-specific diet dependent changes in the microbiota phylotypes[27]. The most likely mechanism, that promotes carbohydrate active enzymes update in the GIT microbiome, is the consumption of foods containing environmental bacteria[28]. It is rational to expect that there are also other factors that affect the configuration of gut microbiota, such as exposure to environmental bacteria and host genetics. All these combined, could result in country-related distinct microbiota profiles considering the different lifestyles, dietary habits and climates of each country. Gut microbiomes sampled from people from Japan, China, Korea and USA revealed profiles clustered according to geographic origin[29]. Analogously, a comparative study between children from Europe and Burkina Faso showed significant country-related differences in the fecal microbiota community[30].

The advent of agricultural societies with the transition from hunting and gathering to agriculture and permanent settlements almost 10,000 years ago must have favored the greatest changes for gut microbial communities. However, as the investigation of fossils has a lot of constraints, other approaches that could shed light in the evolutionary trajectories of the human gut microbiota are being adopted. One of them was recently introduced by a research group that studied the gut microbiota of the Hadza race in Tanzania. The people of this race still live as Paleolithic humans. In other words, they are a modern population of hunter-gatherers. The Hadza gut microbiota is reflecting a functional adaptation to a foraging lifestyle. High bacterial diversity and enrichment of fibrolytic microorganisms seems to be the response to a heavy plant-based diet. Furthermore, the Hadza showed a sex-related divergence in the GM reflecting the sexual division in labor and diet composition. Finally, the absence of Bifidobacterium and the enrichment of potential opportunists, such as Proteobacteria and Spirochaetes, probably demonstrates a different tolerance of their immune system, redefining the notion of what we consider a healthy and an unhealthy gut microbiome structure[31].

NGS based approaches for the study of gut microbiota.

Over the past ten years, the rapid development of next generation sequencing (NGS) technologies increased the number of bases sequenced per run and reduced the sequencing costs. These technological advances drove the development of the novel field of metagenomics. Metagenomics involves the study of the genomes of many organisms simultaneously and is a revolutionary approach in studying complex ecosystems, combining the power of genomics, bioinformatics and systems biology. Thus, provides new access to the microbial world, considering that the vast majority of microbes in nature cannot be grown in the laboratory and be studied with the classical methods of microbiology. Metagenomics overcomes the problems of unculturability and genomic diversity of most microbes.

There are three main NGS strategies which can be used for metagenomics; shotgun sequencing, gene targeted sequencing and metatrascriptomics. Shotgun sequencing is the analysis of the entire microbial community. It is based on the extraction of genomic DNA directly from an environmental sample for the preparation of a NGS library. However, it is important to recognize that the library is not organized in volumes containing the genome of one community member. Instead it consists of millions of clones, each holding a random fragment of DNA. A metagenomic library is like thousands of jigsaw puzzles mixed into a single box. Putting the puzzles together again is a great challenge. The correct assignment of reads into genomes is determined either by mapping the reads in already available reference databases or by *de novo* assembly of the reads into genomes. However, each approach has its own constraints. Targeted sequencing of specific genes enables the study of microbiomes in a more cost-efficient way. In most cases, the 16S rRNA gene is targeted, as it possesses highly conserved regions spaced by hyper-variable regions and can provide a high resolution phylogenetic profile of the community. Using universal primers to amplify the selected locus in a PCR reaction and then sequence the amplified DNA fragments, its easy and fast to perform. However PCR biases can produce ambiguous results[32]. Additionally, recent studies demonstrate that, especially for metagenomic species profiling, the use of many phylogenetic markers is essential[33]. Last but not least, analyzing the entire transcriptome of an environmental site can give a comprehensive view of the community's expression profile, vital for the assessment of different biological processes that characterize the community.

Population genetics and evolution of bacterial species

Although, natural selection operates at the level of the phenotype, its molecular signatures can be revealed, inspecting a population’s genetic variation landscape. In general, when a beneficial allele arise in a population, it tends to increase its frequency within it, reaching fixation at some point if the selective force is strong enough. Conversely, a deleterious allele decreases in frequency till it vanishes from the population. However, in both scenarios, a selected locus can also determine the frequencies of linked but neutral neighboring loci, a phenomenon described as genetic hitchhiking[34,35]. As a result, neutral variants can be dragged to fixation or get lost along with a selected allele, unless recombination breaks down this association. By this means, genetic variation near a positively selected mutation can be greatly reduced, a process known as selective sweep[36]. The extent of a chromosomal region affected by a selective sweep depends on the strength of selection and the local recombination rate, as regions that are distant from the selected site regaining the lost genetic variation due to recombination[37]. Besides, without the effect of recombination all the genetic variation would be lost along with the fixation event.

There is a number of different approaches that can detect the footprint of positive selection on the genome. Some of them are based on the comparison of genetic change between different species and usually are used to detect selection events happened deep in the past (e.g. McDonald-Kreitman test, HKA test, dN/dS, Lewontin-Krakauer test)[38-41], while others investigate selective events within populations and can infer recent activity of selection. For the detection of positive selection in a population, two main strategies have been developed. The first strategy is based on linkage disequilibrium (LD) patterns in the neighborhood of the beneficial mutation[42] and the second strategy is based on the skewing of the allele frequencies (e.g Tajima's D)[43]. LD patterns are generated since independent recombination events on either sides of the selected allele breaks associations between alleles that are located on different sides of the selective sweeps. However, SNPs on the same side still remain partially associated and thus show high levels of LD. The skew of the allelic frequencies occurs since neutral alleles hitchhike with the beneficial allele. Due to recombination however the hitchhiking stops before they reach fixation. Thus, they if they hitchhike they remain in high frequency; if they are not hitchhikers initially, then their frequency decreases. Common to all of these methods, is the use of summary statistics in order to compare the observed data with a null hypothesis of neutrality[44]. Expectations under the null hypothesis can be defined by simulations of a population model that assumes no selection. Under a standard neutral model, the population is reproducing by random mating, the population's size is constant, there is no population subdivision neither generation overlap. In addition, neutrality can be described by even more complicated models and in some cases, empirical estimates of parameters may be available and can be incorporated in the model.

As already mentioned, selective sweeps can alter the frequencies of linked neutral alleles in the population. The distribution of allele frequencies is descibed by the site frequency spectrum (SFS). Thus, skewing of the alleles distribution can be demonstrated in the so called shift of the site frequency spectrum. The SFS is a count of the number of mutations that exist in a frequency of $x_i = i/n$ for $i=1, 2, ... , n-1$, in a sample of size n. Under the standard neutral model, the expected value of x_i is proportional to $1/i$ [45]. In contrast, in case of a selective sweep an excess of low and high frequency derived alleles is observed. However, similar patterns of genetic variation produced by genetic hitchhiking, can be also produced by demographic events. For example, population bottlenecks or expansions may result in an excess of low frequency alleles and many different demographic scenarios are capable for the increase of high frequency alleles as well[46]. Thus, a reliable demographic model could help distinguish the origin of these molecular signatures.

Recently, advances in sequencing technology enabled for population genomics. Therefore, the development of computational tools that can detect positive selection while exploiting the genetic variation patterns across the whole genome was more than a necessity. Kim and Stephan[47] proposed a method that uses composite likelihood ratio (CLR) test, to evaluate the probability of a selective event being responsible for a surplus of derived alleles. Specifically, this test compares the ratio of the composite likelihood of the data under a neutral model against an alternative hypothesis of a complete selective sweep. This statistical test modified has been implemented in computational tools that enable for the detection of sweeps in whole-genome data[48,49]. One important advantage of these tools is that they can either obtain the model for neutrality from the empirical SFS of the entire data (average SFS for all SNPs) or assume a user-defined demographic model for the estimation of SFS. The notion behind is that while the selection's signature is localized around the selected site, demography affects the whole genome and outliers derived from the test procedure are likely sites that a selective sweep has occurred.

In bacteria, patterns of genetic variation depend on the extent to which populations behave clonally. For example in a clonal population every adaptive allele that arises will be perfectly linked to every other allele in the genome. In contrast with eukaryotes, recombination of homologous DNA occurs by gene conversion rather than crossing-over and recombination is decoupled from reproduction. Thus, linkage between nearby loci is expected to be higher than linkage between distant loci in relation to sexually recombining genomes[50]. Distinguishing adaptive mutations within a population depends on the balance between the opposing forces of positive selection, that purges diversity as a new allele approaches fixation and recombination that maintains diversity by unlinking distant regions of the genome from a selective sweep. This balance can be demonstrated by the probability that a polymorphism has arisen due to mutation or recombination. The *r:m* ratio varies considerably among bacteria but is larger than 1, suggesting that recombination is strong in relation to mutation in many species[51]. On the other hand, the mutation rate is universally low in bacteria (10-10 per site per generation)[52].

In this work, we detect selective sweeps in the 66 most abundant bacterial genomes in the human gut analysing metagenomic datasets of 39 healthy individuals. We adopt an SFS-based method for sweep detection implemented by *SweeD.* Additionally, we infer demographic scenarios for each of these bacteria populations, exploiting the genome's estimated SFS, by coalescent simulations with *fastsimcoal2*[53,54].

Methods

Data

All 39 samples used in our analysis, were metagenomic shotgun sequencing data, derived from stool of US healthy individuals, both men and women, in the context of the Human Microbiome Project (http://www.hmpdacc.org/HMASM/). Biological samples were initially sequenced by Illumina WGS plattform and preprocessed by the Human Microbiome Project Consoritum in a way that all human reads were masked, all duplicate reads were removed and low quality bases were trimmed in all reads. All reference genomes were downloaded from the ftp server of NCBI (ftp://ftp.ncbi.nlm.nih.gov/genomes/). For the phylogenetic analysis, we used 16S rRNA sequences derived from stool samples, as part of the HMP.

Data links: ,

Construction of a genomic variation database for human gut microbiome

Metagenomic shotgun sequencing reads from 40 HMP samples were mapped in 66 bacterial genomes of the human gut micriobiome, using *Mosaik aligner* in a way that information of genomic diversity can be retained, with the parameters: *-a all -m all -hs 15 -mmp 0.95 -mmal -minp 0.9 -mhp 100 -act 20*. All genomes downloaded from NCBI ftp server and labeled with a unique tag (BACT_01-BACT66). We created a micriobiome fasta file containing all bacterial genomes, that we used as reference. Only mapped reads retained with samtools and sorted by coordinates with *Picard's SortSam*. All duplicates were marked with *Picard's MarkDuplicate* tool and indexes of bam files were created with *Picard's BuildBamIndex*.

For SNPs identification we used tools from the *GATK* platform. After preparing bam and reference files to be compatible with *GATK* we locally realigned mapped reads, such that the number of mismatching bases is minimized across all the reads, using *RealignerTargetCreator* and *IntelRealigner*. Local realignment serves to transform regions with misalignments due to indels into clean reads containing a consensus indel suitable for standard variant discovery approaches. Unlike most mappers, this process uses the full alignment context to determine whether an appropriate alternate reference (i.e. indel) exists. We omitted the base recalibration step, proposed by *GATK* best practices, as the new estimations for base qualities drastically reduced the number of reads passing the default thresholds for base quality. We suspect that this reduction is due to the absence of a database of known polymorphic sites for our reference genomes. Variant calling was performed with *HaplotypeCaller*, independently for each sample with the parameters: *-ERC GVCF –variant_index_type LINEAR --variant_index_parameter 128000 -ploidy 1 -mbq 20*. This program determines regions of increased variation and builds a De Bruijn-like graph to reassemble the region. Identifying all possible haplotypes in the region, then performs a Smith-Waterman alignment to identify variant sites. In the end, it performs a pairwise alignment of each read against each haplotype using the PairHMM algorithm. This produces a matrix of likelihoods of haplotypes given the read data. These likelihoods are then marginalized to obtain the likelihoods of alleles for each potentially variant site given the read data. The most likely genotype is then assigned to the sample. All genotypeGVCF records produced from variant discovery for each sample separately, were merged into one file, in a way that all possible genotypes found in the cohort for a given site were retained, using *GenotypeGVCFs* with the parameters: *-ploidy 1 -stand_call_conf 30 -stand_emit_conf 10.* SNPs extracted with *SelectVariants* tool.

Estimation of abundances

Following mapping of the metagenomic reads for all samples to reference genomes, we estimated each genome coverage, both in depth and breadth terms. We converted ***bam*** into ***gencov*** files with *bedtools genomecov*. Necessary for this conversion was each genome’s/contig’s length in bps. Lengths were obtained from the respective bam files with a custom script (*getgenomelengths.sh*). We then estimated genome abundances from the genomecov file. Abundances were estimated counting the coverage for each genomic site normalized with the length of the genome and the total number of mapped reads for each sample. We also estimated the percentage of sites that were not covered by a sequencing read for each genome (breadth of coverage). Abundances were estimated with a custom python script (*AbudanceCounter.py*)

Estimation of present genomes in samples

In order to determine the presence or absence of a genome in each biological sample (microbiome), we pooled all breadth of coverage measurements and then performed a cluster analysis. We used k-means algorithm, implemented by an R package, aiming to cluster our data into two groups, one group representative of the present genomes and one for the absent genomes. Thus, we set as threshold for presence in a sample, the mean of the minimum value from the cluster with the higher mean value and the maximum value from the cluster with the lower mean value (m=0.545).

Phylogenetic analysis

Phylogenetic analysis was performed with *RAxML*, a maximum likelihood based algorithm for phylogenetic inference. As input we used 16S rRNA fasta sequences from the reference database of HMP. *RaxML* was called with the parameters below: *-m GTRGAMMA -p 9384503,* where -m specifies the mutation model (here GTR; generalized time reverse mutation model with Gamma parameter estimation for the rate heterogeneity).

Inference of positive selection

All identified SNPs for each bacterial genome in our cohort of individuals were separated and used as input to *SweeD* for the detection of selective sweeps. Data files were converted to the SweepFinder format (SF) which comprises four columns (SNP position, number of derived alleles, number of sequences, and an indication flag whether the SNP is folded or not). Importantly, all genomes found to be absent from a biological sample were not included in our analysis. One SF file was created for all contigs of each genome from which we estimated the entire genome's SFS. This SFS was used downstream in the CLR test step.

Inference of demography

We used *fastsimcoal2*, a coalescent simulation program, to estimate demographic parameters for each bacterial population, based on the SFS estimated for each genome. In short, *fastsimcoal2* simulates the expected SFS under a given set of parameters and computes their (composite) likelihood. *Fastsimcoal2* uses a robust maximization procedure to find those parameters maximizing the composite likelihood. Three type of files were required as input: i) a file containing the observed SFS; ii) a template file (.tpl) in which the evolutionary model to be studied is specified, defining the parameters of interest; iii) an estimation file (.est) in which the parameters' prior distributions are defined. We created all files required as input for the program with a custom python script. (*SFS_ALL_2_obs.py*)

Software links

SweeD

https://www.assembla.com/spaces/sweed/git/source

Fastsimcoal2

http://cmpg.unibe.ch/software/fastsimcoal2/

Results

Analysis of microbiota composition in our cohort

Mapping metagenomic reads in reference genomes may be one of the greatest challenges in metagenomic analysis. First, without having an *a priori* knowledge of the genomes composition in a biological sample, we can only speculate arbitrarily the existence of some selected genomes in it. Second, a metagenomic read can be mapped more than one time in more than one genomes, despite the fact that originates from one genetic locus. At first, we determined the abundances of each bacterial genome in the microbiome of each individual. We found that bacterial genomes are divided into three groups, according to their abundance. One in which genomes are abundant across all individuals, one in which they are sparse and one in which are abundant in some individuals selectively (**Fig.1**). However, the number of reads that are mapped in a genome is not alone necessary and sufficient condition that this genome is present in a sample. There might be some highly conserved regions that can accumulate metagenomic reads even if these reads originate from other genomes in the sample. As a consequence, we also estimated the percentage of the genome that is covered by at least on read, for all genomes included in our analysis. This estimation served as indicator of the presence or absence of a genome in a given microbiome. Clustering analysis pointed out that at least 54,5% of a genome should be covered for a species to be assumed as present in the sample (**Fig.2a,b**). We discovered many species of the phylum Bacteroidetes, such as *Bacteroides vulgatus*, *Bacteroides ovatus*, *Bacteroides sp. 1 1 6*, *Bacteroides uniformis,* to be present in almost all microbiomes in our cohort. Conversely, species *Blautia hansenii*, *Clostridium sp. M62-1*, *Eubacterium cylindroides*, *Holdemania filiformis* and *Prevotella timonensis* were found to be absent in all cases. In addition, some Clostridium species as well as Bacteroides were present only in a few individuals. Surprisingly, E. coli IAI39 strain, a very famous gut microbe, was present in only 6 samples. In general, we ascertained that the existence of a species in our cohort ranged, with the number of samples in which reported as present, covering almost all possible values from 0 to 39. Counts of samples in that a species was assumed to be present are demonstrated in **Fig.2c**. All genomes found to be absent in a given sample were not included in the downstream analysis.

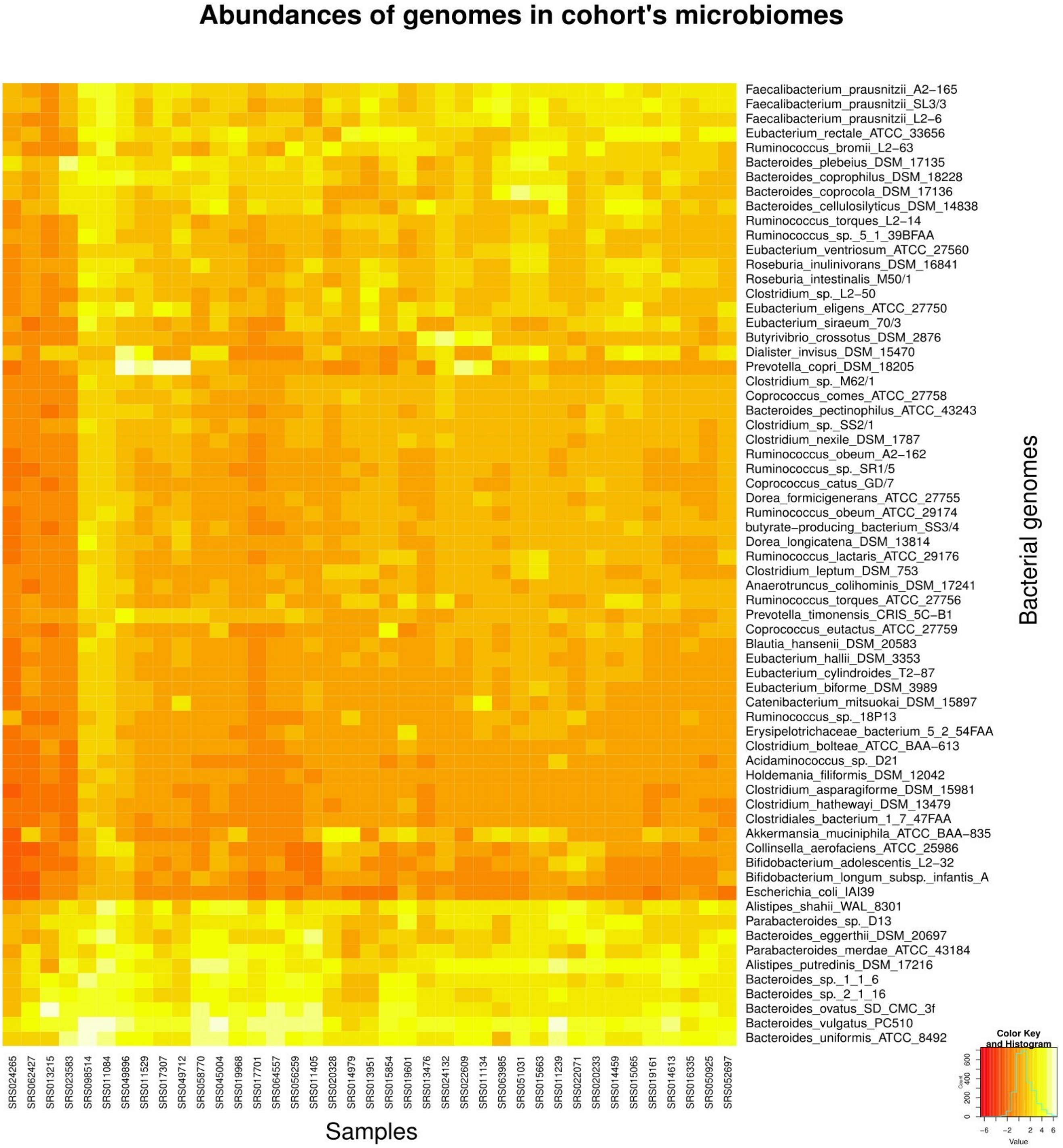

Figure 1 *Heatmap of abundances of 66 bacterial genomes, previously reported as themost abundant in the human gut, among a cohort of 39 healthy individuals. The y axis corresponds to different bacterial species and x axis to different metagenomic samples. Abundances values are in log scale. Light yellow shades denote high abundance values, while dark red shades denote low abundance values.*

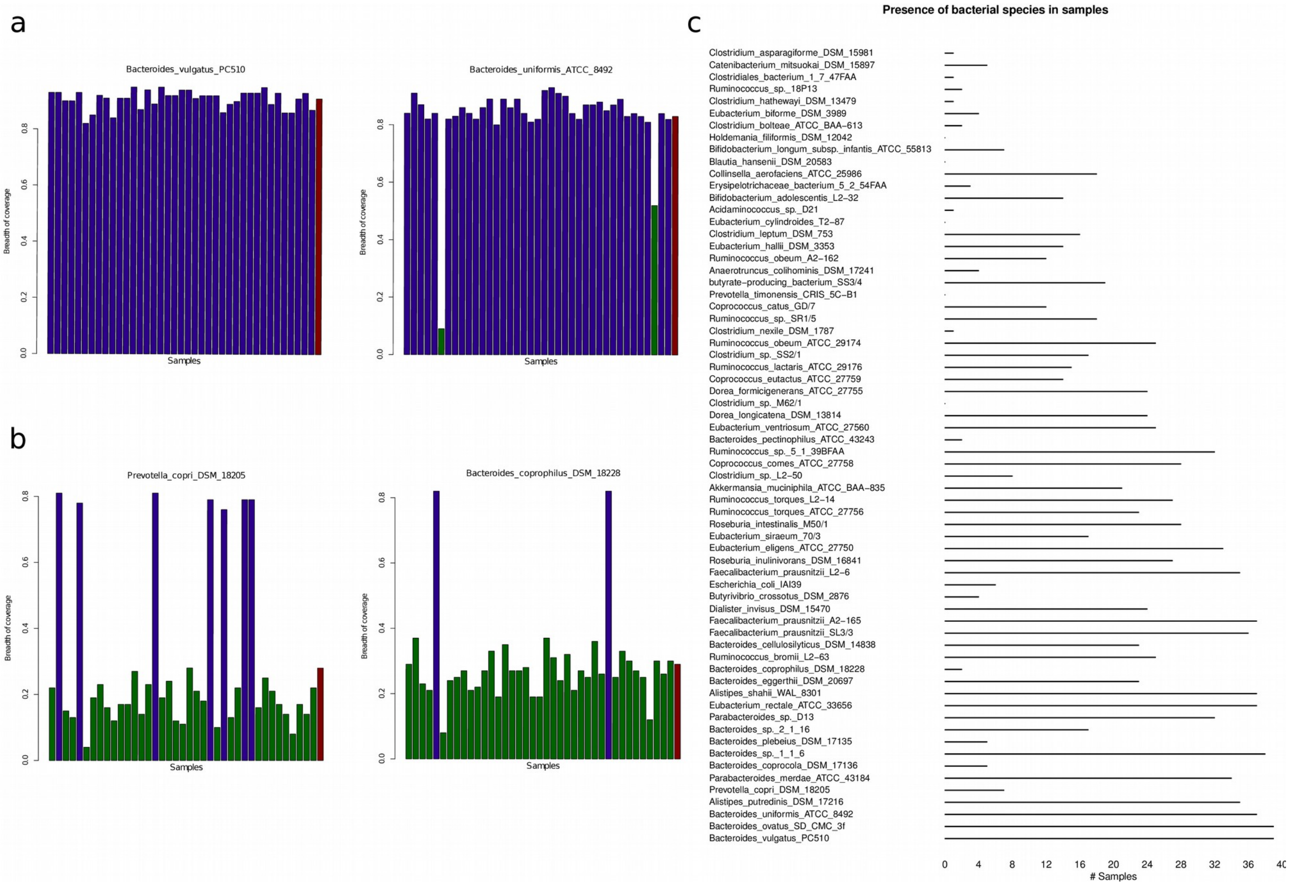

Figure 2 *Clustering of genomes according to the percentage of the genome that is covered by a metagenomic read (Breadth of coverage). Breadth of coverage measurements across all metagenomic samples for* ***a)*** *species found to be present in the majority of cases* ***b)*** *species found to be present in a few cases. Blue color indicates presence, green color indicates absence. Red bars show the average of all measurements.* ***c)*** *The number of metagenomic samples in which a bacterial species estimated to be present.*

Phylogenetically close species tend to have more similar genomes. This implies that a large portion of metagenomic reads that originate from one species can be also mapped in a closely related one. However, it is not clear up to which extend such a fashion exists for bacterial genomes, as there are reported cases in which strains of the same species share less than 40% of their genome. Furthermore, there is mounting evidence that taxonomic units broader than individual species indeed have ecological meaning and, thus, show similar patterns of selection. We constructed a phylogenetic tree of the species included in our analysis, based on 16S rRNA sequences derived from metagenomic stool samples, so that we can explore how phylogenetic distance contributes to the mapping's procedure "species artifacts" and if selection patterns could recognized on a higher taxonimic level. (**Fig.3**)

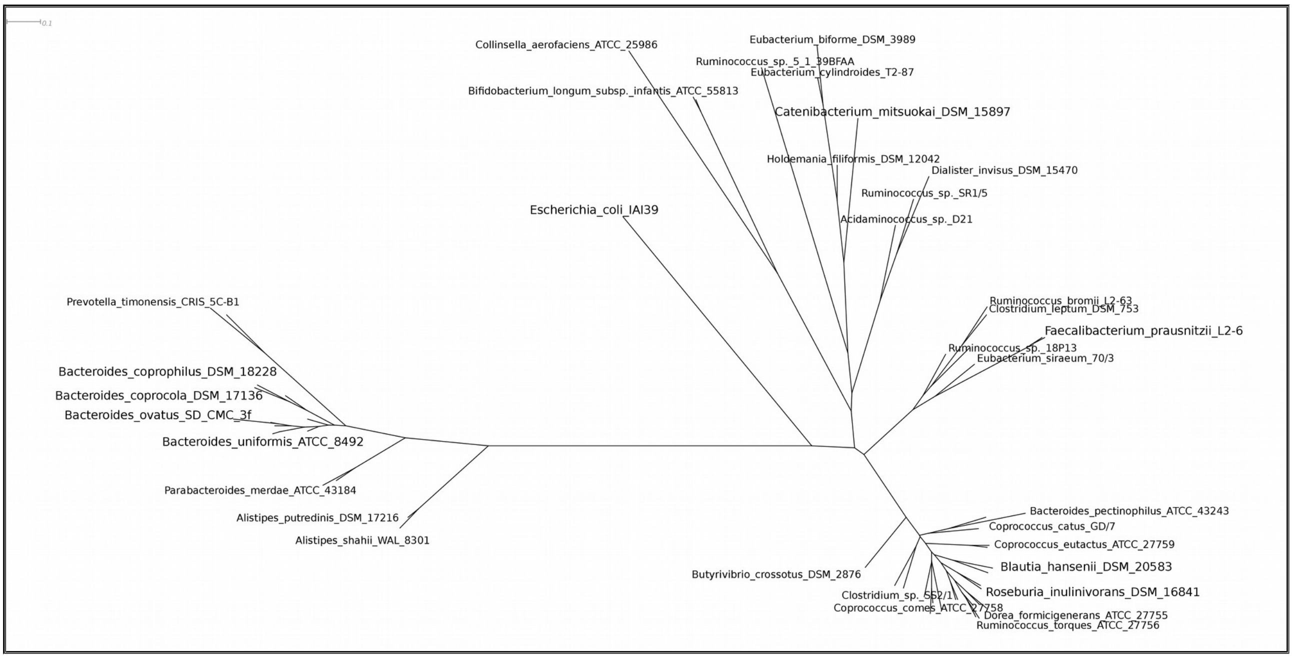
0.1
Collinsella_aerofaciens_ATCC_25986
Bifidobacterium_longum_subsp._infantis_ATCC_55813
Eubacterium_biforme_DSM_3989
Ruminococcus_sp._5_1_39BFAA
Eubacterium_cylindroides_T2-87
Catenibacterium_mitsuokai_DSM_15897
Holdemania_filiformis_DSM_12042
Dialister_invisus_DSM_15470
Ruminococcus_sp._SR1/5
Acidaminococcus_sp._D21
Escherichia_coli_IAI39
Prevotella_timonensis_CRIS_5C-B1
Bacteroides_coprophilus_DSM_18228
Bacteroides_coprocola_DSM_17136
Bacteroides_ovatus_SD_CMC_3f
Bacteroides_uniformis_ATCC_8492
Parabacteroides_merdae_ATCC_43184
Alistipes_putredinis_DSM_17216
Alistipes_shahii_WAL_8301
Ruminococcus_bromii_L2-63
Clostridium_leptum_DSM_753
Faecalibacterium_prausnitzii_L2-6
Ruminococcus_sp._18P13
Eubacterium_siraeum_70/3
Bacteroides_pectinophilus_ATCC_43243
Coprococcus_catus_GD/7
Coprococcus_eutactus_ATCC_27759
Blautia_hansenii_DSM_20583
Butyrivibrio_crossotus_DSM_2876
Roseburia_inulinivorans_DSM_16841
Clostridium_sp._SS2/1
Coprococcus_comes_ATCC_27758
Dorea_formicigenerans_ATCC_27755
Ruminococcus_torques_ATCC_27756

Figure 3 *Phylogenetic distances of all 66 bacterial species based on the 16S rRNA sequence. Species in the leftmost group belong to the Phylum Bacteroidetes. Counterclockwise, species belonging to phyla Firmicutes (two consecutive clusters), Actinobacteria and Proteobacteria are demonstrated. (Not all 66 species are shown in the figure)*

Genomic variation landscape of microbiomes

One of the main goals of this work was to detect selective sweeps in genomes of the most abundant bacterial species of the human gut. Prerequisite for this kind of analysis, was to obtain polymorphism data from our cohort of individuals. Thus, we constructed the genomic variation landscape of the cohort's simplified microbiomes. A plethora of SNPs was identified for each bacterial species. Overall, approximately 11.1 millions polymorphic sites were detected in all genomes. Estimating the number of polymorphic sites across genomes, we distinguished some genomes that possess a relatively high percentage of polympophic sites. On average, 4.6% of genome found polymorphic. Interestingly, 27% of *Dialister invisus* genome was polymorphic across samples, while only few genomes outreach 10%.(**Fig.4a**). We investigated how abundant among the individuals were the sites in which we identified SNPs. Examples of the abundance distribution of polymorphic sites, for a highly polymorphic and a highly abundant species are demonstrated in **Fig.4b**. As we expected, the number of alleles in polymorphic sites was positively correlated with the overall abundance of genome in the cohort.

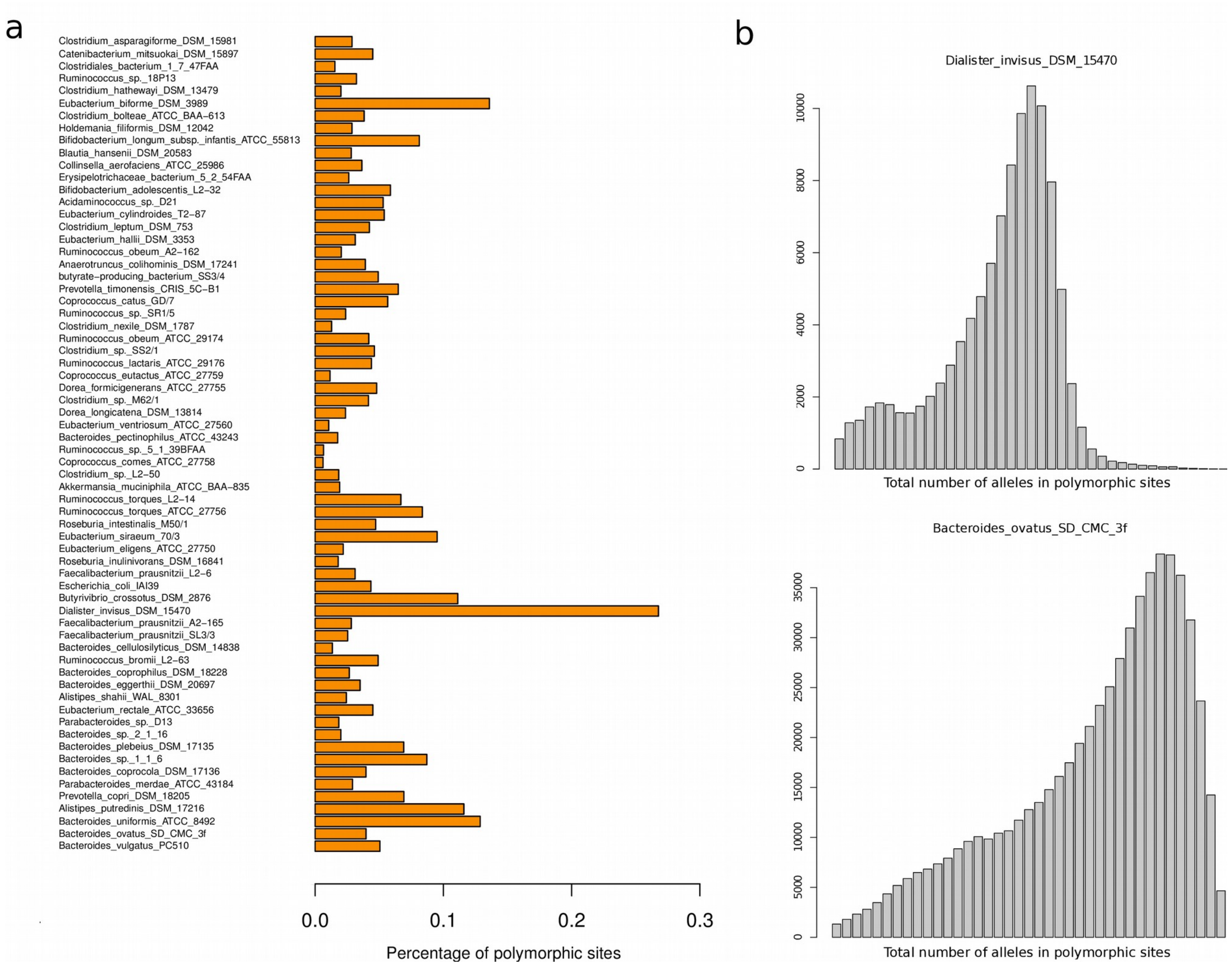

Figure 4 ***a)*** *Percentage of polymorphic sites for all 66 genomes. Mean=4.6%.* ***b)*** *Distribution of the number of alleles in polymorphic sites. Upper plot: Distribution of alleles for Dialister invisus, species high percentage of polymorphic sites. Lower plot: Distribution of alleles for Bacteroides ovatus, highly abundant species.*

Inference of positive selection

The Site Frequency Spectrum in a population under neutrality derives from a hyperbolic distribution. As a consequence, the frequency of alleles in the empirical SFS estimated from an entire genome, should be drawn from this kind of distribution. We used *SweeD* to estimate the empirical SFS for 56 commensal bacterial species of the human gut. (10 species were excluded due to insufficient number of data). We observed different site frequency spectra across species. The expected SFS as described above, was more common among highly abundant species, indicating an impact of missing data in the estimation of SFS. Characteristic of the deviation from the expected was an underestimation of the first class of polymorphisms (singletons). However, the number of polymorphic sites used for the estimation, didn't seem to compensate for the effect of missing data, as in the case of *Dialister invisus* that exhibited a huge number of polymorphic sites. Examples of both expected and deviated SFSs are demonstrated in **Fig.5**.

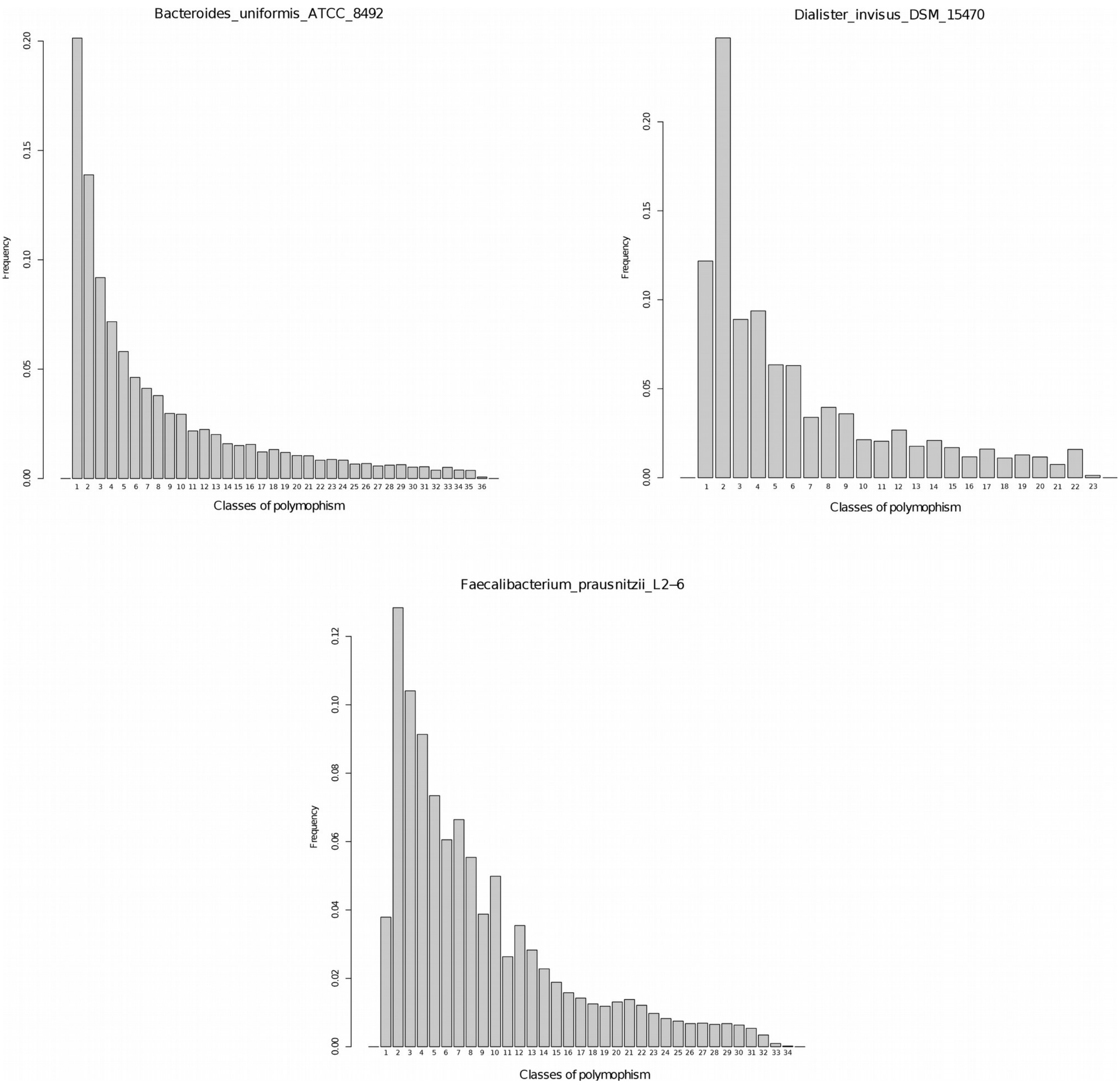

Figure 5 *Distribution of the empirical Site Frequency Spectrum derived from the entire genome of 3 gut commnesal bacterial species. SFS of Bacteroides uniformis, a highly abundant species, has the shape of hyperbolic distribution, as expected. SFS of Dialister invisus, the species with the highest number of polymorphic sites reveals an underestimation of singletons. The same deviation from the expected distribution is obvious in SFS for Faecalibacterium prausnitzii, as well. Both species were abundant in a few individuals in our cohort.*

Inference of demography

The inference of demographic parameters for each of the 56 bacterial populations was based on the SFS estimates. We chose a demographic model that allows for one bottleneck event and one event of population's expansion/shrinkage. Demographic parameters that maximized the composite likelihood of SFS shaped the exact demographic scenario for each population. In **fig6** are summarized all the inferred demographic parameters. In general, a quite common demographic scenario for many bacteria can be described as follows. Back in time, the population's size goes through a bottleneck followed by an expansion that reaches a plateau. Finally, this plateau state is followed by a second event of an ongoing slow reduction in the population size. More exhaustive examination of the different demographic scenarios will reveal more similarities and differences between species history.

GENOMES	NCUR	NANC	NBOT	TBOT	GROWTHRATE	MUTRATE	RECRATE	RESBOT	TENDBOT	MaxEstLhood	MaxObsLhood
Bacteroides_vulgatus_PC510	33373	29021	47536	135649	9.97033E-005	7.94663E-008	1.55932E-009	1.424385	135749	-794077.15	-783259.663
Bacteroides_ovatus_SD_CMC_3f	74077	27551	38386	168178	3.46369E-005	7.8284E-008	2.58153E-008	0.5181905	168278	-1768727.879	-1747305.852
Bacteroides_uniformis_ATCC_8492	48757	60555	74079	115909	3.04534E-005	9.58166E-008	2.54977E-009	1.5193511	116009	-707347.174	-706811.215
Alistipes_putredinis_DSM_17216	56627	39823	53083	88080	2.50675E-005	9.89358E-008	1.31043E-009	0.937415	88180	-354462.975	-353861.351
Prevotella_copri_DSM_18205	16978	62387	36298	114374	0.000256762	0.000000097	7.2156E-009	2.1379432	114474	-284831.276	-279613.506
Parabacteroides_merdae_ATCC_43184	110681	78847	39030	84588	9.75629E-006	9.68767E-008	8.79983E-010	0.3526351	84688	-730510.775	-715250.957
Bacteroides_coprocola_DSM_17136	9185	64962	48540	63163	0.000687966	6.75352E-008	4.26156E-008	5.2847033	63263	-127740.162	-123225.373
Bacteroides_sp._1_1_6	32928	45939	26851	168160	0.000100799	9.82269E-008	3.90234E-009	0.8154458	168260	-1679491.555	-1655887.429
Bacteroides_plebeius_DSM_17135	21502	34658	44300	161717	0.000171508	9.77526E-008	1.68756E-008	2.0602735	161817	-327454.47	-327445.23
Bacteroides_sp._2_1_16	16288	58529	25497	74516	0.000182792	9.76193E-008	1.69211E-008	1.5653856	74616	-451182.455	-445720.297
Parabacteroides_sp._D13	17595	44578	56405	100088	0.000140674	9.74758E-008	2.97333E-008	3.2057403	100188	-647556.329	-644505.41
Eubacterium_rectale_ATCC_33656	58625	42515	78502	121948	1.71441E-005	0.000000095	8.77057E-010	1.3390533	122048	-527523.249	-524696.659
Alistipes_shahii_WAL_8301	85120	44064	40351	156954	2.54312E-005	7.6828E-008	1.77479E-009	0.4740484	157054	-709483.322	-708087.15
Bacteroides_eggerthii_DSM_20697	16011	47715	69765	166584	0.000129458	0.000000096	1.40135E-010	4.3573168	166684	-928342.349	-875810.621
Bacteroides_coprophilus_DSM_18228	103733	12029	62027	6350	-6.28841E-006	4.02321E-008	1.65808E-009	0.5979486	6450	-7588.357	-7588.357
Ruminococcus_bromii_L2-63	139771	34700	65939	164853	4.15245E-006	7.68475E-008	1.31641E-009	0.4717645	164953	-393783.026	-387552.093
Bacteroides_cellulosilyticus_DSM_14838	39968	34811	70460	167954	0.000095124	7.55833E-008	4.87178E-010	1.7629103	168054	-1211385.24	-1182008.899
Faecalibacterium_prausnitzii_SL3/3	32455	33504	32684	168169	0.00010413	7.43442E-008	9.43733E-010	1.0070559	168269	-762580.777	-745391.618
Faecalibacterium_prausnitzii_A2-165	18483	24219	62450	168169	0.000172842	7.48772E-008	2.06504E-009	3.3787805	168269	-600147.327	-588288.07
Dialister_invisus_DSM_15470	42452	30715	82985	163217	5.26597E-005	7.46053E-008	2.45362E-008	1.954796	163317	-262862.352	-259701.506
Butyrivibrio_crossotus_DSM_2876	4763	52857	43428	14425	0.000860001	3.24125E-008	2.23825E-008	9.1177829	14525	-9060.706	-9046.201
Escherichia_coli_IAI39	4480	27717	20973	35633	0.0011072	2.56154E-008	1.13456E-008	4.6814732	35733	-43494.308	-39930.597
Faecalibacterium_prausnitzii_L2-6	16304	67861	52521	167111	0.000222868	7.25416E-008	1.16453E-008	3.2213567	167211	-746110.933	-730949.374
Roseburia_inulinivorans_DSM_16841	14878	46703	75815	127057	0.000215746	7.64636E-008	0.000000005	5.095779	127157	-484489.156	-474176.54
Eubacterium_eligens_ATCC_27750	51713	74180	62362	167695	0.000028782	0.000000096	5.84228E-010	1.205925	167795	-564217.761	-553803.789
Eubacterium_siraeum_70/3	22532	25888	57310	114818	0.000191453	7.77985E-008	1.40061E-009	2.5434937	114918	-332396.009	-331222.37
Roseburia_intestinalis_M50/1	13727	41612	30627	109606	0.00023946	7.54174E-008	8.56108E-010	2.2311503	109706	-453052.591	-441043.266
Ruminococcus_torques_ATCC_27756	4135	36634	64323	36626	0.000786463	7.76777E-008	1.93556E-009	15.5557437	36726	-110620.597	-106419.259
Ruminococcus_torques_L2-14	6963	43859	62130	82815	0.000445542	7.48194E-008	4.04368E-010	8.9228781	82915	-262277.335	-254645.879
Akkermansia_muciniphila_ATCC_BAA-835	191774	46190	56812	168165	-7.29516E-006	7.52963E-008	4.19573E-009	0.2962445	168265	-455181.36	-446817.093
Clostridium_sp._L2-50	8635	80834	62472	92239	0.00065284	0.000000099	5.18139E-008	7.2347423	92339	-227991.964	-217087.638
Coprococcus_comes_ATCC_27758	8616	13734	84678	82043	0.000362363	0.000000074	1.41205E-009	9.8279944	82143	-257124.45	-250558.284
Ruminococcus_sp._5_1_39BFAA	7281	60354	77982	85413	0.000472569	7.93874E-008	7.31152E-009	10.710342	85513	-366033.194	-349265.716
Bacteroides_pectinophilus_ATCC_43243	158736	43121	73196	15927	-1.16941E-005	7.24657E-008	1.56703E-009	0.4611178	16027	-20642.812	-20642.812
Eubacterium_ventriosum_ATCC_27560	6713	19019	32706	57680	0.000553309	9.81876E-008	7.37099E-009	4.8720393	57780	-231825.286	-218198.937
Dorea_longicatena_DSM_13814	6875	21308	78047	75474	0.000476584	7.62588E-008	1.92955E-008	11.3522909	75574	-209981.541	-196010.051
Dorea_formicigenerans_ATCC_27755	5667	112450	62461	69559	0.000529755	7.61738E-008	2.16765E-008	11.0218811	69659	-200477.48	-194134.406
Coprococcus_eutactus_ATCC_27759	10289	34993	33388	162089	0.000210327	9.91803E-008	1.46399E-008	3.245019	162189	-362745.617	-340952.882
Ruminococcus_lactaris_ATCC_29176	6762	48318	62831	59637	0.000526838	7.68305E-008	8.47761E-010	9.2917776	59737	-145249.269	-140693.048
Clostridium_sp._SS2/1	5412	57821	30420	58614	0.000715868	5.78464E-008	4.10149E-010	5.6208426	58714	-140357.828	-129755.933
Ruminococcus_obeum_ATCC_29174	5268	42019	35567	70859	0.000596846	9.71785E-008	4.50113E-010	6.7515186	70959	-306169.149	-277286.879
Ruminococcus_sp._SR1/5	6581	66481	43057	84586	0.000481242	0.000000075	2.56583E-010	6.5426227	84686	-243506.444	-230219.074
Coprococcus_catus_GD/7	3504	69348	37717	41592	0.000990948	9.95702E-008	0.000000012	10.763984	41692	-146340.445	-140604.87
butyrate-producing_bacterium_SS3/4	6810	83607	44705	82365	0.000501664	9.13189E-008	0.000000002	6.5646109	82465	-307938.178	-300247.148
Anaerotruncus_colihominis_DSM_17241	10226	20360	68046	17415	0.0012291	9.61624E-008	1.32566E-009	6.6542147	17515	-53405.365	-52933.625
Ruminococcus_obeum_A2-162	6614	45025	63044	80456	0.000468738	6.92104E-008	3.61374E-010	9.531902	80556	-191224.264	-180806.331
Eubacterium_hallii_DSM_3353	5957	25356	59735	71848	0.000621379	4.68743E-008	5.34402E-009	10.0276985	71948	-132715.049	-123951.45
Clostridium_leptum_DSM_753	7804	44621	68349	59357	0.000458336	7.52375E-008	3.18749E-009	8.7582009	59457	-175915.506	-170263.37
Bifidobacterium_adolescentis_L2-32	7169	64460	30597	47269	0.000595283	6.34421E-008	8.42816E-009	4.2679593	47369	-95886.822	-87965.732
Erysipelotrichaceae_bacterium_5_2_54FAA	10134	40982	29293	14495	0.000145873	3.70703E-008	0.000000043	2.8905664	14595	-11044.023	-11044.023
Collinsella_aerofaciens_ATCC_25986	17549	45064	25271	147520	0.000239137	0.000000078	2.35143E-008	1.4400251	147620	-350879.448	-326787.573
Bifidobacterium_longum_subsp._infantis_ATCC_55813	5817	26438	38065	30535	0.0011015	4.25309E-008	1.57311E-009	6.5437511	30635	-33911.502	-32103.884
Clostridium_bolteae_ATCC_BAA-613	49914	26750	23995	9880	-2.42919E-006	3.91063E-008	3.49006E-009	0.4807269	9980	-16634.735	-16634.735
Eubacterium_biforme_DSM_3989	5900	9572	30030	34422	0.000502801	6.45774E-008	4.11108E-010	5.0898305	34522	-26952.635	-25126.87
Ruminococcus_sp._18P13	261945	35390	66440	25091	-5.31755E-005	7.41607E-008	3.51625E-010	0.253641	25191	-26390.247	-26390.247
Catenibacterium_mitsuokai_DSM_15897	5529	6058	17239	82035	0.00065042	7.28996E-008	7.63313E-009	3.1179237	82135	-79672.484	-72250.614

Figure 6 *Table summarizing the inferred demographic parameters for 56 commensal bacterial species of the huma gut. NCUR: current population size, NANC: ancestral population size, NBOT: population size after bottlneck, TBOT: time in generations since bottlneck, MUTRATE: mutational rate, GROWTHRATE: rate of populations growth. Positive values indicate population shrinkage(backwards in time), RECRATE: recombination rate, RESBOT: , TENDBOT: time in generation since the start of bottleneck event, MaxEstLhood: likelihood estimated from simulations based on SFS. MaxObsLhood: the obtained likelihood by using the observed SFS as the expected SFS when computing the likelihood*

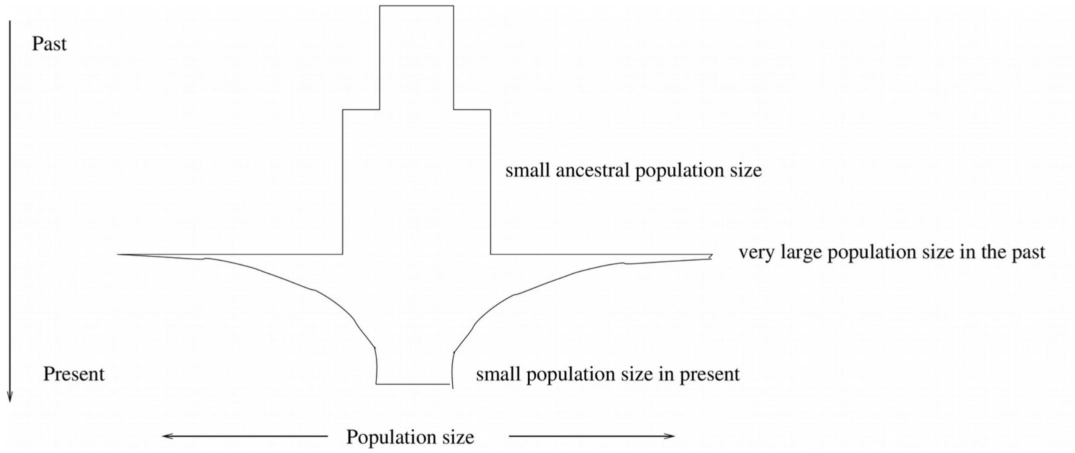
Past
small ancestral population size
very large population size in the past
Present
small population size in present
Population size

***Figure 7**: Illustration of the demographic model for the bacterium Bacteroides vulgatus. In the farthest past, the population size of the bacterium was comparable to the present-day population size. However, about 135k generations in the past there was a very steep increase of the population size, which then started to reduce exponentially until the present-day.*

Discussion

As already mentioned one of the biggest challenges in the science of metagenomics is the procedure of assigning sequencing reads to practically unknown genomes of a biological sample. In our study, we assumed as reference, 66 bacterial genomes previously reported as abundant in the human gut. However, post-hoc analysis revealed that these genomes are not present in all 39 cases tested. This indicates that a pipeline for metagenomic analysis should examine for the presence of selected reference genomes in a biological sample, after the assignment of reads from a metagenomic library to those references, so that misleading conclusions are avoided. Due to the nature of the project, that is the analysis of the genetic variation, we didn’t go through a stringent way of alignment, not to lose some important information regarding the genetic variation in a genetic locus. An alignment step implemented in a way that only uniquely mapped reads are retained could also validate the presence of genomes but would then exclude all metagenomic reads with sequences shared by more than one species. On the other hand, it is rational that an approach that allows a read to be mapped several times in a library results in false positive hits. Indeed, inspection of the genome's coverage distribution revealed genomic regions with peaks of genome coverage. (data not shown). It is plausible that the existence of regions with shared variation could have an impact on our analysis. Alternatively, we could skip all those regions or analyze them separately. Nevertheless, distinguishing between the original genetic variation of a genetic locus from that caused by mapping artifacts can be a very hard task. For this reason, we simplified things assigning to each genomic site, the most prominent genotype for each sample.

The ratio of non-synonymous to synonymous substitution rates has been widely used in genome-wide scans for positive selection in bacteria, often providing evidence for function or gene specific selection. Yet d_N:d_S is inappropriate when comparing either very distantly related strains (because dS becomes saturated with multiple substitutions) or very closely related strains, within which dN:dS is inflated by segregating nonsynonymous polymorphism[54]. Here, we aimed to infer positive selection based on the SFS derived from a "population" of metagenomic libraries. We estimated SFS for 56 bacterial species of the human gut micriobiome. The estimation of SFS from datasets in which part of the information is unknown raised some important concerns. First of all, in order to estimate the allele frequencies for each polymorphic site in the dataset, we have as input two observations; the number of derived alleles and the overall number of all alleles for this particular site. However, for a dataset of N samples, the overall number of alleles n can be a number from 1 to N. In cases that n < N, due to no reported alleles for some samples, we claim that the number of derived alleles observed in our dataset could be different in a theoretical scenario that we had an observation for all N samples. Thus, we estimate the number of derived alleles based on the observation of derived alleles, the observation of overall number of alleles and the known number of all samples, sampling from a hypergeometric distribution. We have evidence that this estimation procedure has some biases and we believe that such biases are the main reason for the underestimation of singletons in many SFSs. One source of bias is the fact that analyzing polymorphic data, all observations of no derived alleles are excluded from the estimation of SFS, because they are in principal non-polymorphic sites. As a result, the assumption that missing data can be drawn from a hypergeometric distribution is wrong. On the contrary, the distribution of missing data is unknown. Thus, the way that the missing observations are handled must be reevaluated.

Last but not least, we assume that all individuals in our cohort are not related to each other. In other words, we claim that the rate of transmission of a bacterial strain from one individual to another is negligible. We also don't take into account horizontal gene transfer between same or different species, a phenomenon quite common for bacteria. Accounting for HGT may have a great impact on the estimation of SFS under neutrality. HGT can be modeled as a gene conversion event and has already been mathematically incorporated in coalescent theory[55].

References

1. Gill, S. R. et al. (2006). Metagenomic analysis of the human distal gut microbiome. *Science* **312**:1355–1359.

2. The Human Microbiome Project Consortium. (2013). Structure, Function and Diversity of the Healthy Human Microbiome. *Nature* **486**:207–214.

3. O'Hara, A. M., and Shanahan, F. (2006). The gut flora as a forgotten organ. *EMBO* Rep. **7**:688–693.

4. Dominguez-Bello MG, Costello EK, Contreras M, Magris M, Hidalgo G, Fierer N, et al. (2010) Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body

habitats in newborns. *Proc Natl Acad Sci USA* **107**:11971-11975.

5. Koenig, J. E., Spor, A., Scalfone, N., Fricker, A. D., Stombaugh, J., Knight, R. et al. (2011). Succession of microbial consortia in the developing infant gut microbiome. *Proc Natl Acad Sci USA* ***108*** **Suppl 1:**4578–4585.

6. Qin, J., Li, R., Raes, J., Arumugam, M. et al. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464(7285)**:59–65.

7. Caporaso JG, Lauber CL, Costello EK, et al. (2011). Moving pictures of the human microbiome. *Genome Biol* **12**:R50.

8. Eckburg, P. B. et al. (2005) Diversity of the human intestinal microbial flora. *Science* **308**:1635–1638.
9. Schloissnig, S., Arumugam, M., Sunagawa, S., Mitreva, M. et al. (2013). Genomic variation landscape of the human gut microbiome. *Nature* **493(7430)**:45–50.

10. Koropatkin, N. M., Cameron, E. A., and Martens, E. C. (2012). How glycan metabolism shapes the human gut microbiota. *Nat. Rev. Microbiol.* **10**:323–335.

11. Kamada, N., Seo, S.-U., Chen,G.Y.,andNúñez, G. (2013). Role of the gut microbiota in immunity and inflammatory disease. *Nat. Rev. Immunol.* **13**:321–335.

12. Cryan, J. F., and Dinan, T. G. (2012). Mind-altering microorganisms: the impact of the gut microbiota on brain and behaviour. *Nat. Rev. Neurosci.* **13**:701–712.

13. Collins, S. M., Surette, M., and Bercik, P. (2012). The interplay between the intestinal microbiota and the brain. *Nat. Rev. Microbiol.* **10**:735–742.

14. Mayer, E., Tillisch, K., & Gupta, A. (2015). Gut / brain axis and the microbiota. *J. Clin. Invest.* **37**:49–62.

15. Russell, W. R., Hoyles, L., Flint, H. J., and Dumas, M.-E. (2013). Colonic bacterial metabolites and human health. *Curr. Opin. Microbiol.* **16**:246–254.

16. Rombeau, J. L. & Kripke, S. A. (1990). Metabolic and intestinal effects of short-chain fatty acids. J*PEN. J. Parenter. Enteral Nutr.* **14**: S181–S185.

17. Ninonuevo, M. R. et al. (2006). A strategy for annotating the human milk glycome. *J. Agric. Food Chem.* **54**:7471–7480.

18. Chaturvedi, P., Warren, C. D., Buescher, C. R., Pickering, L. K. & Newburg, D. S. (2001). Survival of human milk oligosaccharides in the intestine of infants. *Adv. Exp. Med. Biol.* **501**:315–323.

19. Fuhrer, A. et al. (2010). Milk sialyllactose influences colitis in mice through selective intestinal bacterial colonization. *J. Exp. Med.* **207**:2843–2854.

20. Fagarasan, S., Kawamoto, S., Kanagawa, O. & Suzuki, K. (2010). Adaptive immune regulation in the gut: T cell-dependent and T cell-independent IgA synthesis. *Annu. Rev. Immunol.* **28**:243–273.

21. Littman, D. R. & Rudensky, A. Y. (2010). Th17 and regulatory T cells in mediating and restraining inflammation. *Cell* **140**:845–858.

22. Sudo, N. et al. (2004). Postnatal microbial colonization programs the hypothalamic–pituitary–adrenal system for stress response in mice. *J. Physiol.* **558**:263–275.

23. Larauche, M., Mulak, A. & Tache, Y. (2012). Stress and visceral pain: from animal models to

clinical therapies. *Exp. Neurol.* **233**: 49–67.

24. Tannock, G. W. & Savage, D. C. (1974). Influences of dietary and environmental stress on microbial populations in the murine gastrointestinal tract. *Infect. Immun.* **9**:591–598.

25. David, L. A., Maurice, C. F., Carmody, R. N., Gootenberg, D. B., But- ton, J. E., Wolfe, B. E., et al. (2014). Diet rapidly and reproducibly alters the human gut microbiome. *Nature* **505**:559–563.

26. Candela, M., Biagi, E., Maccaferri, S., Turroni, S., and Brigidi, P. (2012). Intestinal microbiota is a plastic factor responding to environmental changes. *Trends Microbiol.* **20**:385–391.

27. Walker, A.W. et al. (2011). Dominant and diet-responsive groups of bacteria within the human colonic microbiota. *ISME J.* **5**:220–230.

28. Hehemann, J.H. et al. (2010). Transfer of carbohydrate-active enzymes from marine bacteria to Japanese gut microbiota. *Nature* **464**:908–912.

29. Nam, Y.D. et al. (2011). Comparative analysis of Korean human gut microbiota by barcoded pyrosequencing. *PLoS ONE* **6**:e22109.

30. De Filippo, C. et al. (2010). Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa. *Proc. Natl. Acad. Sci. U.S.A.* **107**:14691–14696

31. Schnorr, S. L., Candela, M., Rampelli, S., Centanni, M., Consolandi, C., Basaglia, G., et al. (2014). Gut microbiome of the hadza hunter-gatherers. *Nat. Commun*. **5**:3654.

32. Lee CK, Herbold CW, Polson SW, et al. (2012). Groundtruthing next-gen sequencing for microbial ecology-biases and errors in community structure estimates from PCR amplicon pyrosequencing. *PLoS One* **7**:e44224.

33. Sunagawa, S., Mende, D. R., Zeller, G., Izquierdo-Carrasco, F., Berger, S. a, Kultima, J. R. et al. (2013). Metagenomic species profiling using universal phylogenetic marker genes. *Nature Methods* **10(12)**:1196–9.

34. Maynard Smith, J., and J. Haigh, (1974). The hitchhiking effect of a favourable gene. *Genet. Res*. **23**:23–35

35. Kaplan, N. L., R. R. Hudson and C. H. Langley, (1989). The “hitchhiking effect” revisited. *Genetics* **123**: 887–899

36. John Maynard Smith and John Haigh (1974). The hitch-hiking effect of a favourable gene. Genetical Research, 23, pp 23-35. doi:10.1017/S0016672300014634.

37. Richard R. Hudson, Martin Kreitman and Montserrat Aguadé. (1987). A Test of Neutral Molecular Evolution Based on Nucleotide Data. *Genetics* vol.**116**:1153-1159

38. Hurst LD. (2002). The Ka/Ks ratio: diagnosing the form of sequence evolution. *Trends Genet.* **18(9)**:486

39. Wright SI, Charlesworth B. (2004). The HKA test revisited. *Genetics* **168(2)**:1071–1076

40. McDonald JH, Kreitman M. (1991). Adaptive protein evolution at the Adh locus in Drosophila.

Nature **351(6328)**:652–654

41. Lewontin RC, Krakauer J. (1973). Distribution of gene frequency as a test of theory of selective neutrality of polymorphisms. *Genetics* **74**:175–95

42. Sabeti PC, Reich DE, Higgins JM, Levine HZP, Richter DJ, et al. (2002). Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419(6909)**:832–37.

43. Fu Y-X. (1997). Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics* **147(2)**:915–25.

44. Vitti, J. J., Grossman, S. R., & Sabeti, P. C. (2013). Detecting natural selection in genomic data. Annual Review of *Genetics* **47**:97–120.

45. Nielsen, R. (2005). Molecular signatures of natural selection. Annual Review of *Genetics* **39**:197–218.

46. Jensen, J. D., Kim, Y., DuMont, V. B., Aquadro, C. F., & Bustamante, C. D. (2005). Distinguishing between selective sweeps and demography using DNA polymorphism data. *Genetics* **170**:1401–1410.

47. Kim, Y., and W. Stephan,. (2002) Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics* **160**:765–777.

48. Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, Bustamante C. (2005). Genomic scans for selective sweeps using SNP data. *Genome Res.* **15**:1566–1575.

49. Pavlidis, P., Živković, D., Stamatakis, A., & Alachiotis, N. (2013). SweeD: Likelihood-based detection of selective sweeps in thousands of genomes. *Molecular Biology and Evolution* **30(9)**: 2224–2234.

50. Shapiro, B. J., David, L. a., Friedman, J., & Alm, E. J. (2009). Looking for Darwin’s footprints in the microbial world. *Trends in Microbiology* **17**:196–204.

51. Didelot, X. and Falush, D. (2007). Inference of bacterial microevolution using multilocus sequence data. *Genetics* **175**:1251–1266.

52. Drake, J.W. (1991). A constant rate of spontaneous mutation in DNA- based microbes. *Proc. Natl. Acad. Sci. U. S. A.* **88**:7160–7164.

53. Excoffier, L. and M. Foll. (2011). fastsimcoal: a continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. *Bioinformatics* **27**:1332- 1334.

54. Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., and M. Foll. (2013). Robust demographic inference from genomic and SNP data. *PLOS Genetics* **9(10)**:e1003905.

55. Kryazhimskiy, S. and Plotkin, J.B. (2008). The population genetics of dN/dS. *PLoS Genet.* **4**: e1000304

56. Wiuf, C., & Hein, J. (2000). The coalescent with gene conversion. *Genetics* **155(1):**451–462.