Genomic variation landscape of the human gut microbiome

Siegfried Schloissnig1*, Manimozhiyan Arumugam1*, Shinichi Sunagawa1*, Makedonka Mitreva2, Julien Tap1, Ana Zhu1, Alison Waller1, Daniel R. Mende1, Jens Roat Kultima1, John Martin2, Karthik Kota2, Shamil R. Sunyaev3, George M. Weinstock2 & Peer Bork1,4

Whereas large-scale efforts have rapidly advanced the understanding and practical impact of human genomic variation, the practical impact of variation is largely unexplored in the human microbiome. We therefore developed a framework for metagenomic variation analysis and applied it to 252 faecal metagenomes of 207 individuals from Europe and North America. Using 7.4 billion reads aligned to 101 reference species, we detected 10.3 million single nucleotide polymorphisms (SNPs), 107,991 short insertions/deletions, and 1,051 structural variants. The average ratio of non–synonymous to synonymous polymorphism rates of 0.11 was more variable between gut microbial species than across human hosts. Subjects sampled at varying time intervals exhibited individuality and temporal stability of SNP variation patterns, despite considerable composition changes of their gut microbiota. This indicates that individual–specific strains are not easily replaced and that an individual might have a unique metagenomic genotype, which may be exploitable for personalized diet or drug intake.

With the increasing availability of individual human genomes, various theoretical and practical aspects of genomic variation can be deduced for individuals and the human population as a whole14. Like sequenced human genomes, the number of human gut metagenomes (currently mostly derived from Illumina shotgun sequencing of stool samples) is increasing exponentially. Given the importance of the gut microbiota in human health4–8 and a growing number of studies reporting associations between gut microbiota and diseases5–8, an understanding of genomic variation in gut microbial populations will probably trigger applications towards human well-being and disease.

For example, in the common gut commensal bacterium Escherichia coli, just three point mutations in two genes can confer clinically relevant antibiotic resistance4, and natural variation in a single gene can lead to pathogenic adaptation4. Even within pathogenic species in the gut, closely related coexisting strains can exhibit different pathogenic potentials due to minor genomic variation5. These examples illustrate how genomic variation within gut microbes could confer phenotypes that require personalized care or treatment of the host.

Studies based on 16S ribosomal RNA gene surveys or whole metagenome shotgun sequencing characterized taxonomic and functional compositions of healthy individuals’ and intestinal bowel disease patients’ gut microbiota at the genus or species level9–12. Variation in taxonomic abundance as well as functions encoded by these gut microbiota have been described between individuals9,11 and used to stratify individuals according to their gut community compositions into enterotypes13. However, genomic variation within species, which leads to their phenotypic diversity and adaptations to different environments, has only been studied in a few taxa, such as Citrobacter spp7.

An early landmark study on a small data set described metagenomic variation in an acidic biofilm microbiome of low complexity14. The population structure for one species in that habitat was studied and positive selection was observed in some genes15. Another recent study resolved multiple clinical isolates of methicillin-resistant Staphylococcus aureus and delineated its epidemiology and micro-evolution based on genomic variation16. With the availability of hundreds of deeply sequenced human gut metagenomes9,11,17, sufficient data are becoming available for quantitative analyses of the genetic structure of complex microbial communities, allowing the study of many species at the same time.

Here we analysed 1.56 terabases of sequence data from 252 stool samples from 207 individuals (Supplementary Table 1 and Supplementary Notes) obtained from the MetaHIT project (71 Danish, 39 Spanish; all sampled once11), the NIH Human Microbiome Project (94 US samples; 51 individuals sampled once, 41 sampled twice and 2 sampled three times9), and Washington University (three US samples; all sampled once12). Our goals were to: (1) develop a framework for genomic variation analysis using metagenomic shotgun data; (2) gather basic knowledge on the genomic variation landscape in gut metagenomes; and (3) gain insights into the individuality, temporal stability and biogeography of metagenomic variation.

Framework for metagenomic variation analysis

We used 1,497 prokaryotic genomes to generate a set of reference genomes (Supplementary Table 2) for the analysis of genomic variation in gut microbial species in 252 samples (on average 6.2 ± 4.1 gigabases (Gb) were analysed). Pairwise comparisons of 40 universal marker genes18,19 identified in these genomes were performed to create a set of 929 clusters based on a 95% DNA identity threshold recommended for identifying species30. The genome recruiting the highest number of reads in a cluster was selected as reference for that species (see Methods and Supplementary Information).

Using the same 95% identity threshold, we mapped 7.4 billion metagenomic reads (42% of the total, 91% thereof uniquely) with an average length of 80 base pairs (bp) to the 929 reference genomes (Supplementary Tables 1 and 3). To avoid mapping artefacts (for example, caused by high coverage of prophages), we required ≥40%
of each reference genome to be covered by reads (corresponding to the gene content similarity between two strains of E. coli\(^2\)). The resulting 101 prevalent species with base-pair coverage from 12× to 32,400× (Fig. 1 and Supplementary Fig. 1) were subjected to genomic variation analysis.

To enable comparative analyses in multiple metagenomes and to identify low-frequency variants not detected when analysing them individually, we used multi-sample calling\(^2\) to identify SNPs, short insertions/deletions (indels; 1–50 bp) and structural variations (>50 bp) in each genome, although structural variations were largely underestimated due to small insert sizes (Supplementary Information). We only called variants with allele frequency ≥1% (the conventional definition of polymorphism\(^2\)) and supported by ≥4 reads. False-positive rates were estimated at 0.71% for SNPs and 1.04% for structural variations (Supplementary Information, Supplementary Tables 4 and 5, and Supplementary Fig. 2).

**Genomic variation in prevalent gut species**

We identified 10.3 million SNPs in 101 genomes (3.1% of the total 329 Mb) across 252 samples from 207 subjects, almost as many as the 14.4 million SNPs recently identified in 179 human genomes\(^2\). Within an individual the rate was lower (on average 1.21%, see Supplementary Information, Supplementary Table 6), yet SNPs kb\(^{-1}\) within an individual the rate was lower (on average 1.21%, see Supplementary Information). In each genome, although structural variations were largely underestimated due to small insert sizes (Supplementary Information), we only called variants with allele frequency ≥1% (the conventional definition of polymorphism\(^2\)) and supported by ≥4 reads. False-positive rates were estimated at 0.71% for SNPs and 1.04% for structural variations (Supplementary Information, Supplementary Tables 4 and 5, and Supplementary Fig. 2).

Figure 1 | Genomic variation statistics for 101 gut microbial species prevalent in 252 samples from 207 individuals. Genomic variation statistics were calculated for 101 prevalent gut microbial species, operationally defined as having ≥10× cumulative (over all samples) base-pair coverage with at least one sample exhibiting a genome coverage of ≥40%. The 66 dominant species (indicated by an asterisk), which account for 99% of the mapped reads, were used for analyses that required high base-pair coverage. Species names are given without strain specifications unless this would result in duplicate entries. The blue point cloud plots show the coverages (≥1×) in all samples, with the blue dot above indicating the cumulative coverage and the red dot the maximum coverage across all samples. Grey shaded areas indicate the level of base-pair coverage at which abundance effects have only minor effects on SNPs kb\(^{-1}\) and pN/pS ratios of the pooled samples (Supplementary Information). SNP counts appear to saturate at approximately 500×, with minor increases at higher coverages probably due to the sampling of rare variants at low rates. In individual samples, pN/pS is largely stable from a coverage of 10× onward (Supplementary Fig. 7), corresponding to approximately 200× cumulative coverage in our sample set. Nucleotide diversity π follows SNPs kb\(^{-1}\) closely, as does the derived measure of pπ(N)/pπ(S) with respect to pN/pS.

We annotated the genes of the prevalent genomes using orthologous groups from eggNOG\(^2\) (Supplementary Information) and found that the orthologous groups with the highest SNP density were enriched in functions related to conjugal transfer of antibiotic resistance (Supplementary Table 7). For example, the orthologous group with the highest average SNP density across samples was the clindamycin resistance transfer factor btgA (NOG119724), required for conjugative transmission of plasmids. Mutations commonly accrued from the process of conjugation may account for increased diversity among conjugation-associated functions\(^2\). Additionally, CRISPR-associated proteins, responsible for conferring resistance in bacteria, were also found among the orthologous groups with high SNP densities (Supplementary Information and Supplementary Table 7).

The large number of SNPs provided the opportunity to compare, for the first time at such scale, the evolution of different coexisting species across a large cohort of individuals. To evaluate selective constraints in these species in their natural habitat, we estimated the ratio of non-synonymous to synonymous polymorphism rates\(^2,25,26\) (pN/pS) within each species in every sample (Fig. 1 and Supplementary Information). pN/pS characterizes selective constraint at the level of a population contrary to the more commonly used dN/dS that characterizes it between individual species\(^2,26\). To validate pN/pS ratios, we estimated genetic variation using the sample-size-independent nucleotide diversity π, and found that π is highly correlated with SNPs kb\(^{-1}\) (Fig. 1 and Supplementary Fig. 5). The derived measures of π(N)/π(S) and π(non-degenerate sites)/π(fourfold degenerate sites), the latter of...
which is less dependent on specific properties of mutation spectra such as transition and transversion ratios, were coherent with pN/pS (Supplementary Fig. 6).

The pN/pS ratio of a genome within a sample remained stable at coverages higher than 10× (Supplementary Fig. 7)—yet another indication of few false SNP calls—and was on average 0.11, but varied considerably for different species (0.04 to 0.58) in accordance with dN/dS ratios estimated independently in a number of interspecies comparisons between closely related bacteria and archaea27,28.

pN/pS across gut species and individuals

Because meaningful comparison of genomic variation requires both breadth (across samples) and depth (in number of base pairs) of sequencing, we focused on the 66 most dominant species that attracted >99% of the reads (Fig. 1). Their relatively low pN/pS ratios were constant across different hosts (Fig. 2a and Supplementary Table 8), which may indicate similar selective constraints across individuals. Thus, the evolution of gut species is probably dominated by long-term purifying selection and drift rather than rapid adaptations to specific host environments. The wide range of these ratios across species may suggest that different gut species face different evolutionary constraints.

To investigate how different gut species respond to the pressure from the gut environment, we compared the pN/pS ratios of individual genes in Roseburia intestinalis and Eubacterium eligens, which differed considerably in their overall mean pN/pS ratios (0.236 (R. intestinalis) versus 0.131 (E. eligens) from 106 and 147 samples, respectively) despite having comparable average base-pair coverages (Supplementary Information). Whereas 75% of the genes in R. intestinalis had systematically higher pN/pS ratios compared to their orthologues in E. eligens, few others showed considerable deviations (Fig. 2b and Supplementary Table 9), indicative of differing evolutionary constraints for these genes. For example, galK, the gene encoding galactokinase, an essential enzyme in the Leloir pathway for galactose metabolism in most organisms29, was among the lowest in terms of pN/pS ratio in R. intestinalis but among the highest in E. eligens (0.03 and 0.48, respectively; Fig. 2b, c). Although present in E. eligens, this gene may not exert its main function (see also ref. 30), as E. eligens cannot ferment galactose, nor the galactose-containing disaccharides lactose and melibiose31. On the other hand, R. intestinalis is known to ferment melibiose32, indicating that its galK gene is functional (Supplementary Information). Thus, the same gene can be under tight negative selection in one species but under more relaxed negative selection in another.

Our framework allowed us additionally to obtain information on all genes in each sample (Supplementary Information). As expected, we found that housekeeping genes had usually lower pN/pS ratios. For example, the DNA-dependent RNA polymerase β-subunit gene was consistently among the genes exhibiting the lowest pN/pS ratios across samples and species (Supplementary Table 10). Less obvious examples included genes related to type IV secretion systems used to transfer DNA between microbes33 and involved in host interactions of both pathogenic34 and commensal bacteria35, specifically in anti-inflammatory responses and immune modulation35. The low pN/pS ratio of genes related to type IV secretion systems suggests that maintaining genome plasticity and antibiotic resistance through conjugative transposition is essential in the constantly changing environment of the gut and that the interaction with the host immune system is under purifying selection (Supplementary Table 10). We also found a few conserved unknown, but apparently gut-microbe-specific, proteins that exhibited low pN/pS ratios, suggesting that they perform important yet hitherto unexplored functions (Supplementary Table 11).

Among the genes or orthologous groups with consistently the highest pN/pS ratios were many transposases and antimicrobial resistance genes including the gut-specific gene bile salt hydrolase (BSH)36 (Supplementary Table 10). Conjugated bile acids (CBAs) secreted by the hosts repress microbial growth and upregulate the host mucosal defence system. BSHs are involved in the initial reaction in the metabolism of CBAs by gut microbes37. Their high pN/pS ratio may be indicative of the genomic plasticity necessary to metabolize and survive the variety of different bile acids present in the gut37.

Temporal stability of individual SNP patterns

Several studies on adult human gut microbial samples from a few individuals have suggested that within-individual differences over time are smaller compared to between-individual differences in microbial species composition and abundance38–40. Within a larger cohort, individuality of host-associated microbiota has been reported on the basis of 16S rRNA gene profiling of fingertip-associated communities41, whereas other studies on a few samples have investigated the persistence of specific strains over time42,43. However, intra-species variation at nucleotide resolution at the whole-genome level and accompanying changes in species abundances within the human gut over long time periods (>1 year) have not been studied yet in large cohorts. It is unclear if the concept of resident strains is common to other prevalent species, if host-specific strains are retained over time, and how fast they evolve inside the gut environment.

To explore these questions, we used 88 gut metagenomes from 43 healthy US subjects (a subset of our cohort) from whom at least two samples were obtained at different time points with no antibiotics treatment in between (Supplementary Table 12). To measure how similar the subpopulations (strains) of the dominant species were between two samples, we estimated the fixation indices (FST) between the populations (Supplementary Information). Because this measure depends on allele frequencies, which cannot be determined accurately at low base-pair coverage, we also estimated the fraction of alleles shared between the samples out of all polymorphic sites (only 49 genomes that accrued 40% genome coverage in at least two samples were used and genomes with >10× base-pair coverage were down-sampled to 10×; Supplementary Information and Supplementary Fig. 3). Because the fraction of shared SNPs depends on the number of variable sites, we developed a heuristic allele sharing similarity score that takes into account both the number of variable sites and the fraction of shared alleles (Supplementary Information).

When we compared all 252 samples, FST was significantly lower and allele sharing significantly stronger between different samples from the same individuals than between samples from different individuals (Mann–Whitney U-test: P < 0.001 for both; see Fig. 3a, b and Supplementary Information). The same trend was observed, albeit much weaker, based on species compositions (Fig. 3c), in line with previous observations from microbial composition-based results38–40. Intra-individual variation being smaller than inter-individual variation does not require that samples from the same individual are more similar to each other than to any other sample in the tested cohort. Our results showed for both measures of variation similarity that all but one of the 88 multi-time-point samples had the highest similarity to another sample from the same individual, which was not true when comparing species abundance over time (Fig. 3c and Supplementary Information). This indicates that species abundance in gut microbiota cannot serve as a fingerprint of an individual whereas variation patterns might.

We also tested whether differences in FST and allele sharing decreased over time, which may indicate a divergence of the strains or a horizontal transmission of strains from the environment; however, the individual-specific variation patterns remained stable over all of the time intervals monitored (Fig. 3). Although this stability should be verified for longer periods as well as when antibiotic treatment or other gut microbiota-challenging events have taken place, our observation indicates that healthy human individuals retain specific strains (see also Supplementary Tables 12 and 13) for at least 1 year.

In contrast to the strong evidence for individuality and temporal stability of SNP patterns, we did not observe a significant geographical
Figure 2 | pN/pS ratios of 66 dominant species reveal more variation between species than between individuals. a, A heat map of pN/pS ratios for the 66 dominant species (rows) and 207 individuals (columns; only the first time-point per individual) is shown and summarized by species (box plots on the right). Rows and columns are ordered by their mean pN/pS ratios, which vary considerably between species, but have a tighter bandwidth across samples. Two genomes that are exceptions to this trend (indicated by an asterisk) might indicate higher strain diversity. The panel above the heat map indicates the continent of residence for each individual. A significant difference was found in the mean pN/pS ratios between the two continents, although this is probably an effect of lower sequencing depths of European samples (see, for example, top-right corner). b, The distributions of average pN/pS ratios of individual genomes from Roseburia intestinalis and Eubacterium eligens (both highlighted in a) illustrate that, although base-pair coverages are similar, the pN/pS ratio of R. intestinalis is higher in general. The relative pN/pS ratios of orthologous groups in the two species are shown in the inset, with the average log₂ ratio indicated by the solid line and the random expectation by the dashed line. Outliers can be revealed this way, like the galactokinase gene (galK), the pN/pS of which is among the lowest in R. intestinalis (bottom panel). The cumulative read coverage is shown in grey with synonymous (green) and non-synonymous (brown) changes marked at the nucleotide positions they occur.

Discussion
We have established a framework for gut microbial genomic variation analysis using metagenomic data and identified in a single analysis—involved 252 stool samples from 207 human individuals—almost as many SNPs in the human gut microbiome as the 1000 Genomes Project recorded in 179 human individuals over several years. The stable pN/pS ratios of gut microbial species across individuals suggest that host conditions (such as diet, genetic differences, and immune...
Shannon distance13 (counts and the ratio of shared SNPs into account (different individual (blue dots) in terms of population similarity that takes allele variation patterns were remarkably stable over time (Fig. 3a, b), which compared to constraints common to the human population (such as geographical separation, diseases, host-genetic and life-style/diet factors) shape the distribution of gut microbial strains and segregating SNPs within the population. The absence of clear geographical stratification implies that stable differences in variation patterns of gut species are not explained by large-scale structures of local microbial populations. They may rather be a result of genetic drift due to population bottlenecks that could occur not only during the colonization of the infant gut but also by processes causing community shifts during adult life stages, followed by a rapid population growth accompanied by purifying selection. This model suggests that the source of genetic variation in human gut microbial populations is less likely to be new mutations within the host than the variation in the initial colonizing populations or transmissions from the environment. This would imply that most allelic variants analysed in this study segregate at timescales greatly exceeding human generation time.

The introduction of large-scale variation analysis in metagenomic data of complex communities and the discovery of individual metagenomic variation profiles open up several applications. It is now possible to screen in silico for many pathogenic or antibiotics resistance variants in the population. Once a sample has been analysed, the data can also be used in the future given the temporal stability of SNP profiles. As it took years to identify marker genes and variations for diseases or phenotypes in the human genome, the variation landscape uncovered here can only be seen as the beginning to find molecular biomarkers including particular variants that reveal useful information for human health and well-being.

**METHODS SUMMARY**

**Mapping to non-redundant genomes.** A reference genome set representing 929 species was derived from a total of 1,497 published prokaryotic genomes, based on a median sequence identity of 95% in 40 universal, single-copy marker genes18-20. Metagenomic reads from 252 samples were aligned to these 929 genomes on a median sequence identity of 95% in 40 universal, single-copy marker genes18-20. Metagenomic reads from 252 samples were aligned to these 929 genomes using the same 95% sequence identity cutoff. Coverage. For each genome, we calculated the sample-specific base-pair coverage and the number of bases of the genome covered by at least one read. For a genome to be considered we required a cumulative depth of coverage of $\geq 10^6 \times$ across all samples. To remove species that are not present in our cohort, yet attract reads due to highly conserved regions, we required at least 40% breadth of the genome coverage (the criterion for the species to be considered present) from at least one sample. Variation detection. We performed SNP calling on the pooled samples and only considered bases with a quality score $\geq 15$. We required SNPs to be supported by $\geq 4$ reads and to occur with a frequency of $\geq 1%$. Structural variations were detected using Pindel48. False-positive rates in SNP and structural variation detection were estimated using nonsense and frameshift mutations in 40 essential single copy marker genes.

$\pi$ and $F_{ST}$. We estimated nucleotide diversity ($\pi$) and fixation indices ($F_{ST}$) on the basis of allele frequencies.
pN/pS ratio. SNPs occurring in coding regions were classified as synonymous or non-synonymous. Genes from the non-redundant genomes were annotated using eggNOG orthologous groups allowing calculation of pN/pS ratios at the level of genomes, orthologous groups and genes.

**Pairwise sample comparisons.** Similarity in strain populations between two samples was estimated using a similarity score based on shared SNPs and using FST.

**Received 25 January; accepted 25 October 2012.**

**Published online 5 December 2012.**

1. International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449, 851–861 (2007).

2. The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* 467, 1061–1073 (2010).

3. Backhed, F., Ley, R. E., Sonnenburg, J. L., Peterson, D. A. & Gordon, J. I. Host-bacterial mutualism in the human intestine. *Science* 307, 1915–1920 (2005).

4. Hooper, L. V., Midtvedt, T. & Gordon, J. I. How host-microbial interactions shape the host. *Annu. Rev. Nutr.** 22, 283–307 (2002).

5. Bagel, S., Hüllen, V., Wiedemann, B. & Heisig, P. Impact of evolutionary and ecological factors on metagenomic orthologous groups allowing calculation of pN/pS ratios at the level of microbial genomes. *PLoS Comput. Biol.* 6, e1000875 (2010).

6. Simmons, S. L. Population genomic analysis of strain variation in *Escherichia coli*. *Antimicrob. Agents Chemother.* 45, 2866–2870 (2001).

7. Morowitz, M. J. Microbial diversity and the genetic nature of microbial species. *Proc. Natl Acad. Sci. USA* 106, 587–590 (2006).

8. Tyson, G. W., Wefze, M. & Woese, C. R. Evolution of MRSA during hospital transmission and host-specific adaptation of bacterial species during antibiotic perturbation. *Proc. Natl Acad. Sci. USA* 108, 4554–4561 (2011).

9. Zoetendal, E. G., Akkersmann, A. D. & De Vos, W. M. Temperature gradient gel electrophoresis analysis of 16S rRNA from human fecal samples reveals stability and host-specific communities of active bacteria. *Appl. Environ. Microbiol.* 64, 3854–3859 (1998).

10. Fierer, N. et al. Forensic identification using skin bacterial communities. *Proc. Natl Acad. Sci. USA* 107, 6477–6481 (2010).

11. Tenenail, O., Skurnik, D., Picard, B. & Denamur, E. The population genetics of commensal *Escherichia coli*. *Nature Rev. Microbiol.* 8, 207–217 (2010).

12. Jernberg, C., Lofmark, S., Edlund, C. & Jansson, J. K. Long-term ecological impacts of antibiotic administration on the human intestinal microbiota. *ISME J.* 1, 56–66 (2007).

13. Suzuki, R., Shiota, S. & Yamaoka, Y. Molecular epidemiology, population genetics, and pathogenic role of *Helicobacter pylori*. *Infect. Genet. Evol.* 12, 203–213 (2012).

14. Yamaoka, Y. *Helicobacter pylori* typing as a tool for tracking human migration. *Clin. Microbiol. Infect.* 15, 839–841 (2009).

15. Achtman, M. & Wagner, M. Microbial diversity and the genetic nature of microbial species. *Nature Rev. Microbiol.* 6, 431–440 (2008).

16. Morelli, G. et al. *Yersinia pestis* genome sequencing identifies patterns of global phylogenetic diversity. *Nature Genet.* 42, 1140–1143 (2010).

17. Ye, K., Schulz, M. H., Long, Q., Apweiler, R. & Ning, Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* 25, 2862–2871 (2009).

18. Suzuki, R., Shiota, S. & Yamaoka, Y. Molecular epidemiology, population genetics, and pathogenic role of *Helicobacter pylori*. *Infect. Genet. Evol.* 12, 203–213 (2012).

19. Yamaoka, Y. *Helicobacter pylori* typing as a tool for tracking human migration. *Clin. Microbiol. Infect.* 15, 839–841 (2009).

20. Achtman, M. & Wagner, M. Microbial diversity and the genetic nature of microbial species. *Nature Rev. Microbiol.* 6, 431–440 (2008).

21. Morelli, G. et al. *Yersinia pestis* genome sequencing identifies patterns of global phylogenetic diversity. *Nature Genet.* 42, 1140–1143 (2010).

22. Ye, K., Schulz, M. H., Long, Q., Apweiler, R. & Ning, Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* 25, 2862–2871 (2009).

23. Supplementary Information is available in the online version of the paper.

**Acknowledgements** The authors are grateful to J. Korbel and the members of the Bork group at EMBL for discussions and assistance, especially S. Powell for performing some of the computations. We thank the EMBL IT core facility and Y. Yuan for managing the high-performance computing resources. We would like to thank J. I. Gordon for providing three of the samples used also grateful to the European MetAHT consortium and the NIH Common Fund Human Microbiome Project Consortium for generating and making available the data sets used in this study. The research leading to these results has received funding from EMBL, the European Community’s Seventh Framework Programme via the MetaHIT (HEALTH-F4-2007-201052) and iHMS (HEALTH-F4-2010-261376) grants as well as from the National Institutes of Health grants U54HG030379 and U54HG004968.

**Author Contributions** P.B. and G.M.W. conceived the study, P.B., M.A., G.M.W. and S.R.S. designed the analyses. S.S., Sh.S., M.A., M.M., J.T., A.Z., A.W., D.M., J.R.K., J.M. and K.K. performed the analyses. M.A., Sh.S., S.I.S. and P.B. wrote the manuscript. All authors read and approved the manuscript.

**Author Information** Single nucleotide polymorphism data have been submitted to dbSNP under accession numbers ss539238913–ss549853572. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to P.B. (bork@embl.de) or G.M.W. (gweinste@genomewustl.edu).