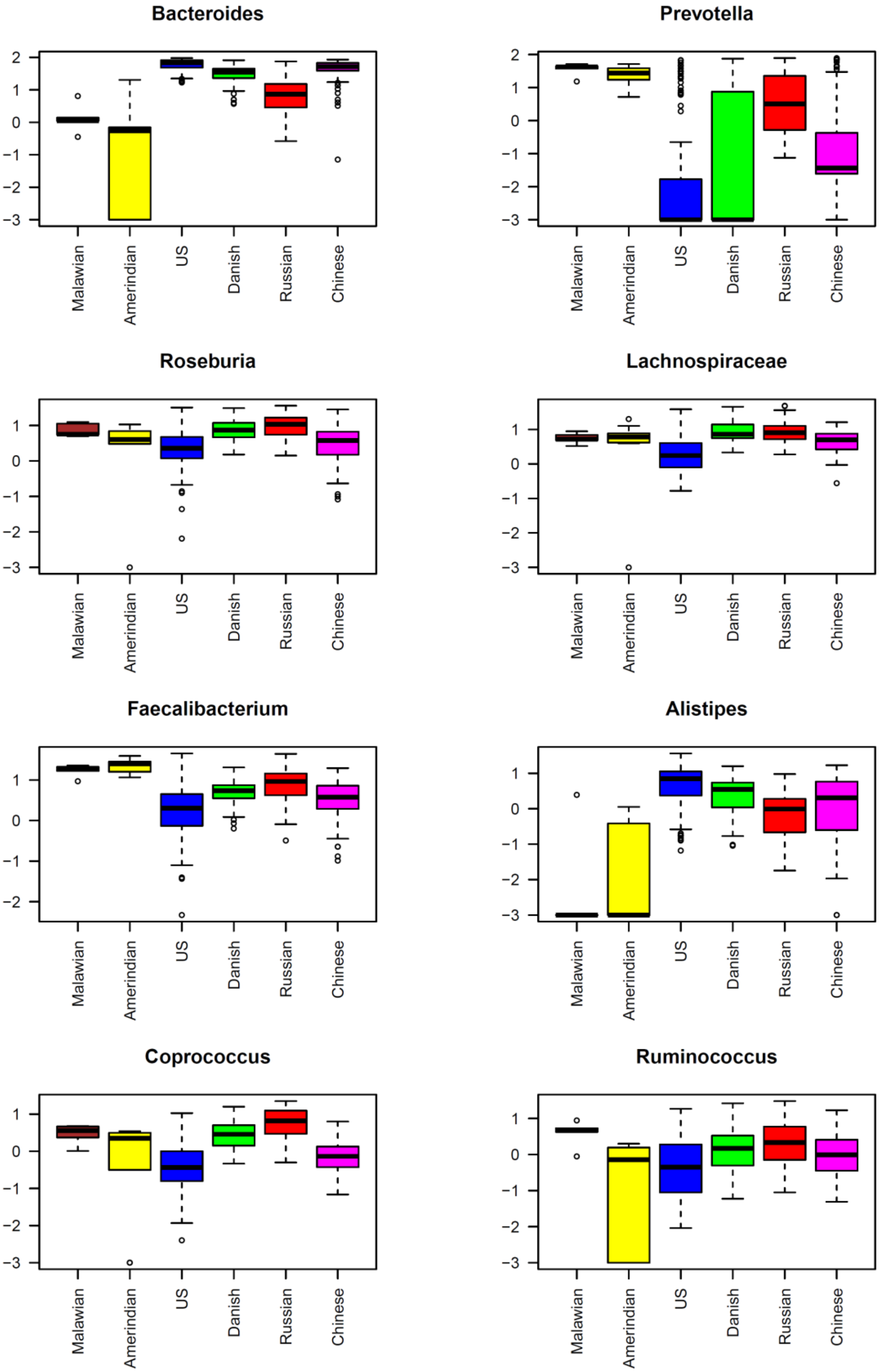
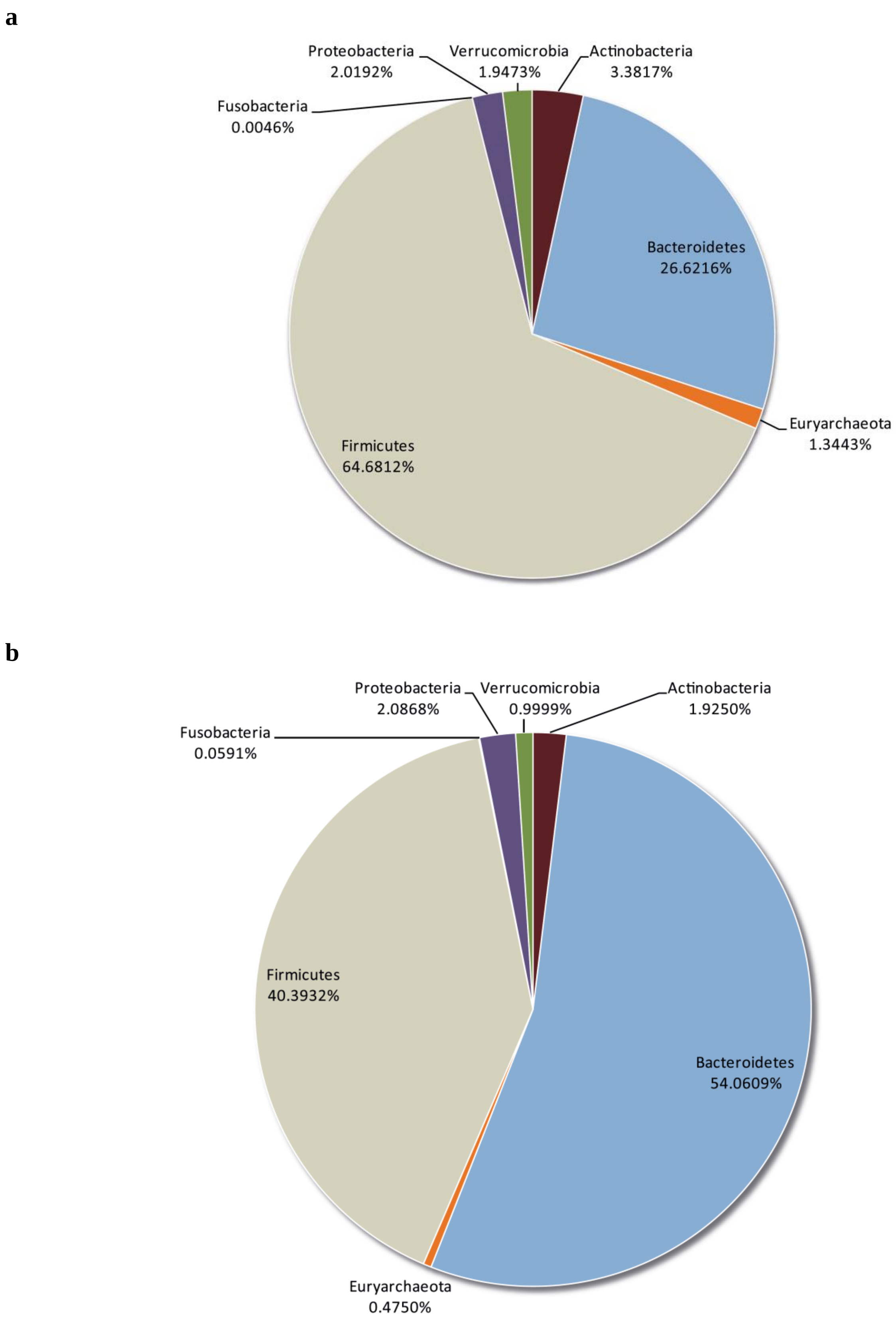
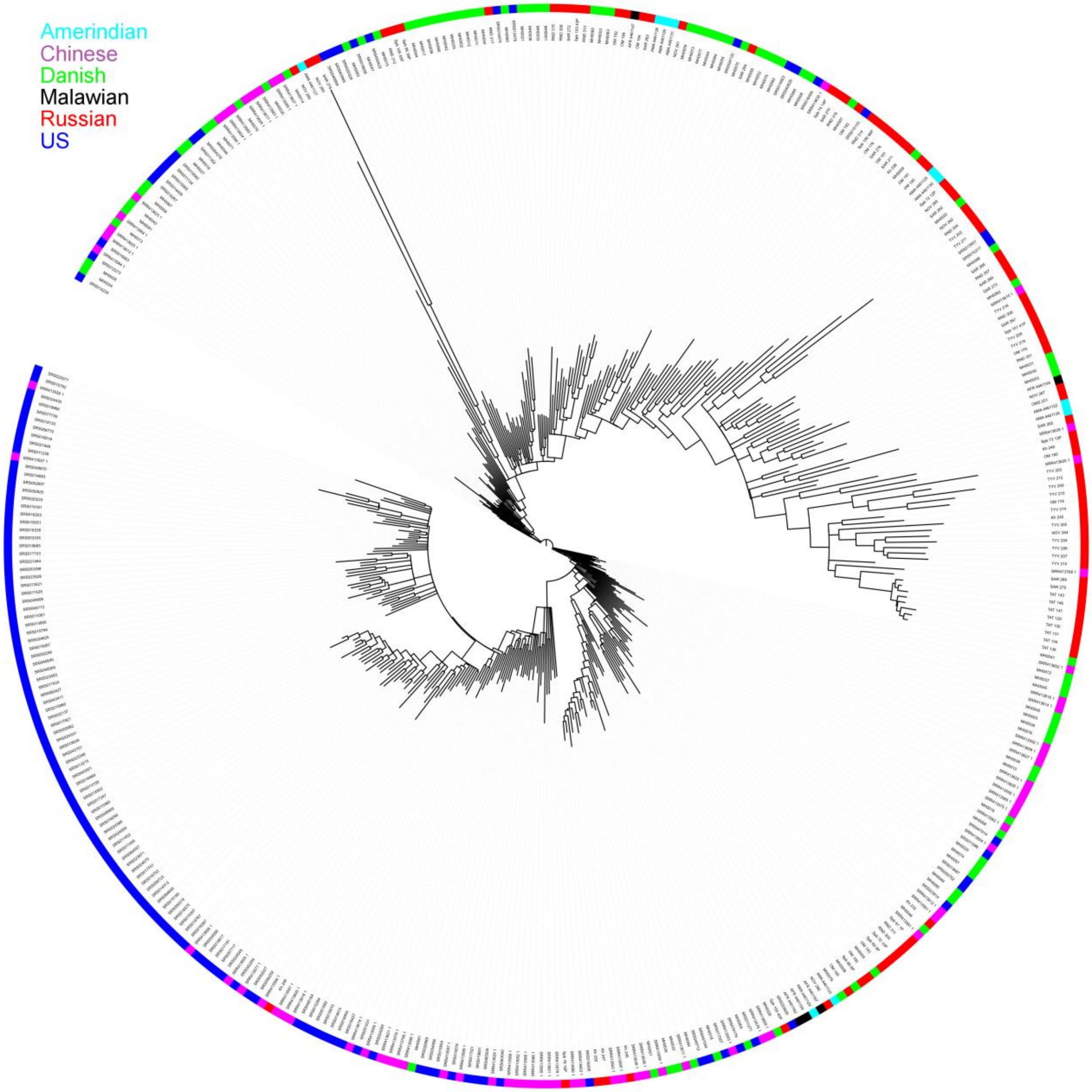
Supplementary Figures



Supplementary Fig. S1 - Nationwide contributions of the most abundant genera.
The figure shows $\log_{10}$ of the relative percentage of genera, forming 80% of total abundance. (Russian (*n*=96), Danish (*n*=85), US (*n*=137), Amerindian (*n*=10) Malawian (*n*=5) and Chinese (*n*=70)). Bottom and top of the boxes denote the 1^{st} and the 3^{rd} quartile, the band inside boxes is median and whiskers are the lowest datum still within 1.5 IQR of the lower quartile, and the highest datum still within 1.5 IQR of the upper quartile.

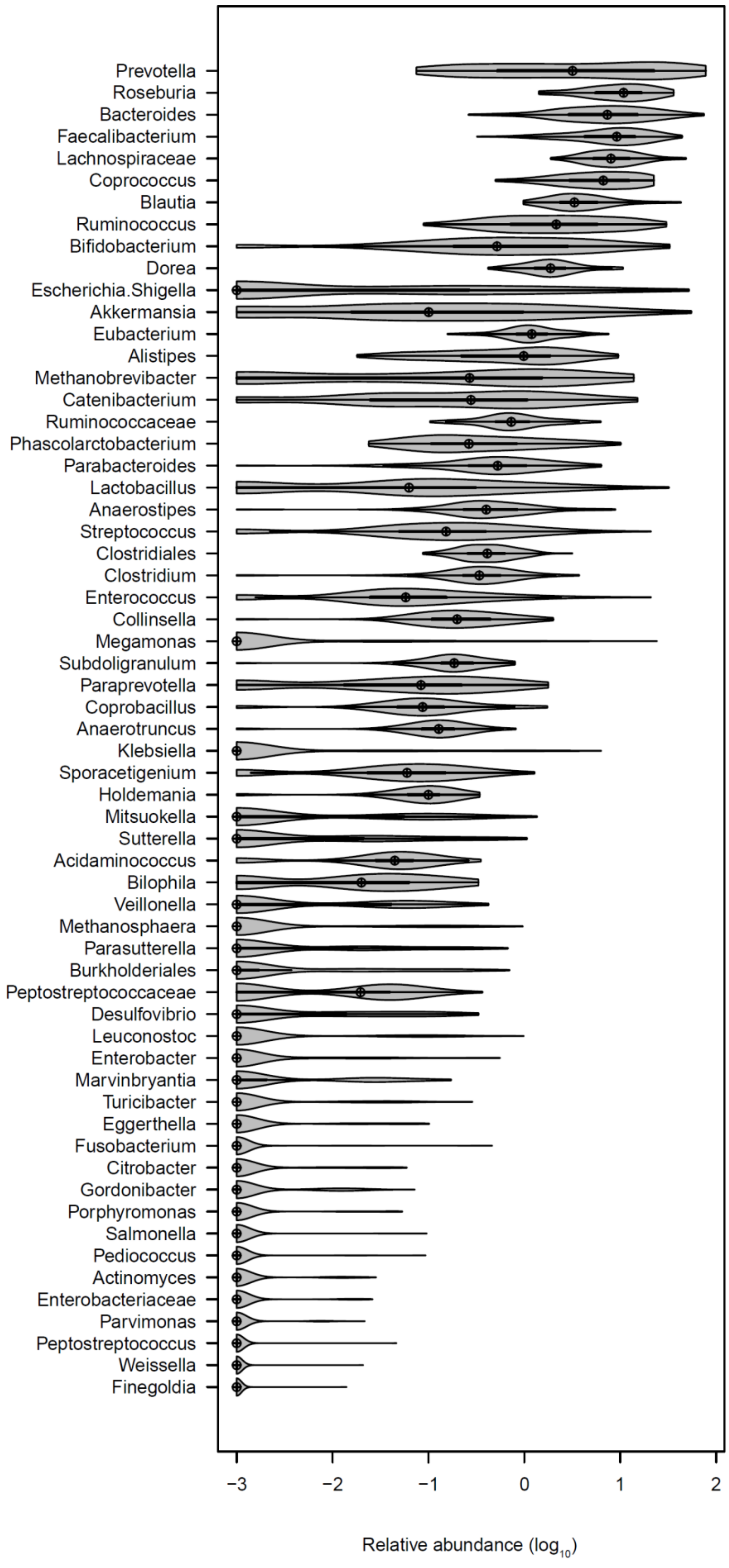


Supplementary Fig. S2 - Phylum-level taxonomic composition. a, Russian samples (*n*=96). **b**, non-Russian samples (*n*=307).

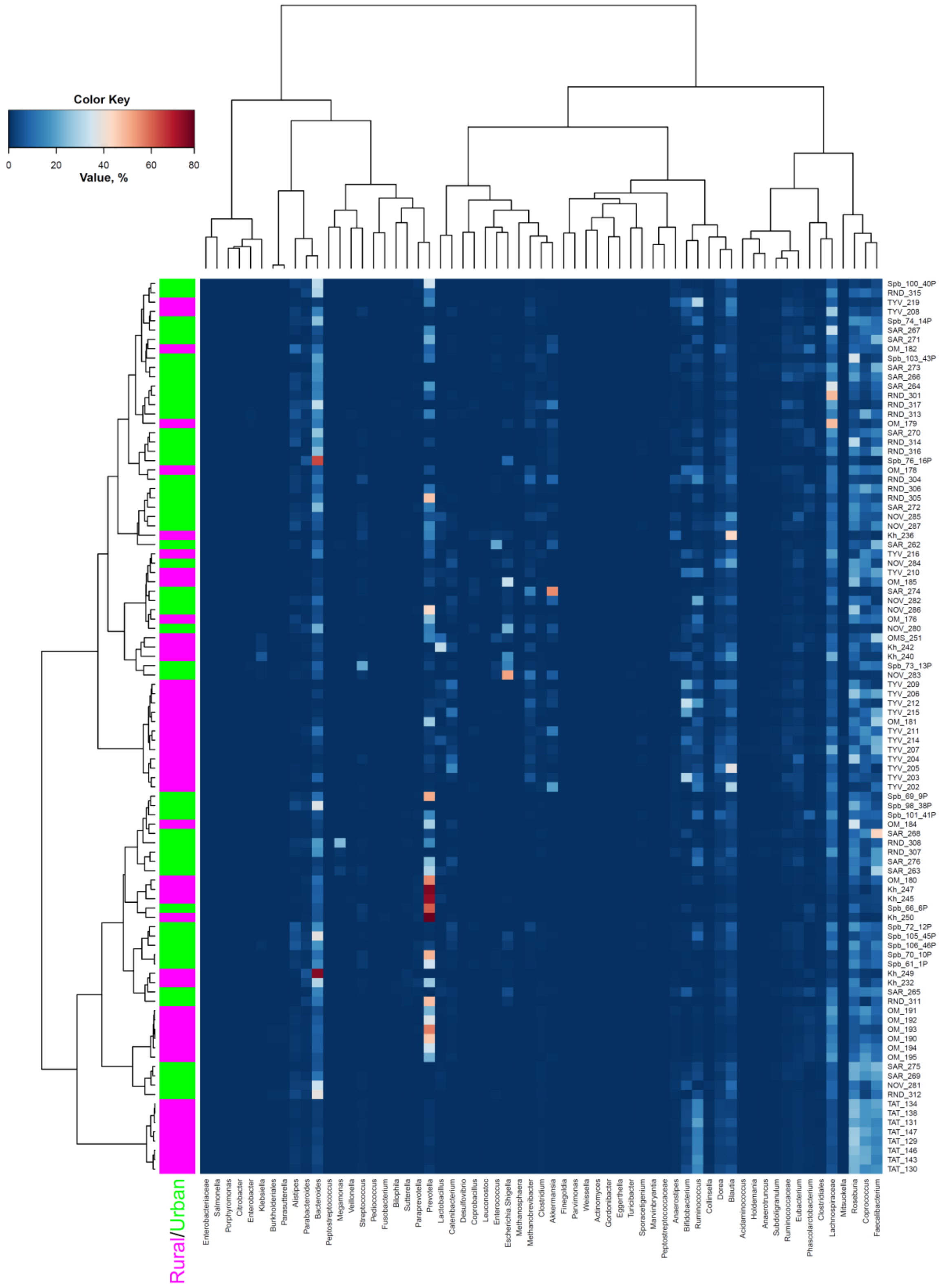


Supplementary Fig. S3 – Global diversity of the human gut metagenomic composition (visualised in iTOL).

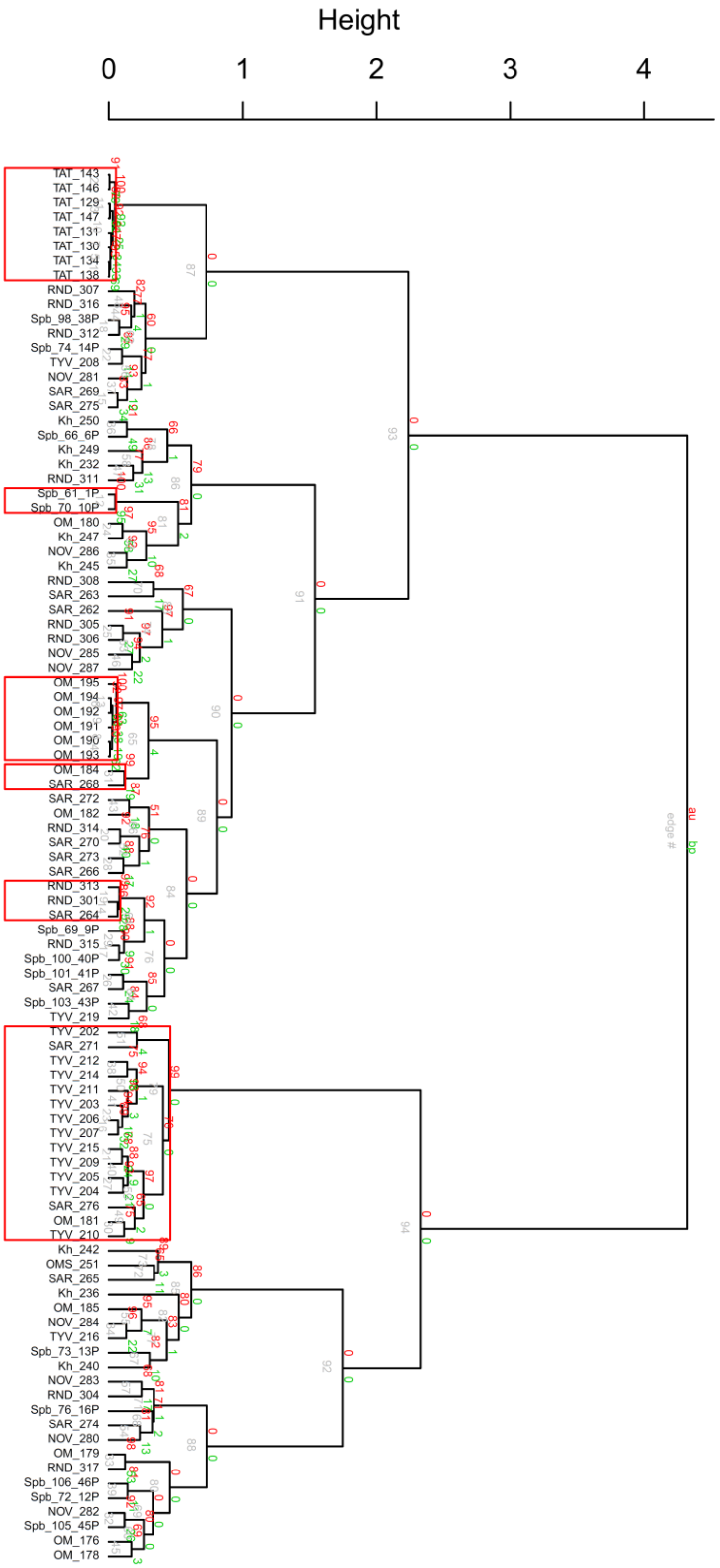
a.


Prevotella
Roseburia
Bacteroides
Faecalibacterium
Lachnospiraceae
Coprococcus
Blautia
Ruminococcus
Bifidobacterium
Dorea
Escherichia.Shigella
Akkermansia
Eubacterium
Alistipes
Methanobrevibacter
Catenibacterium
Ruminococcaceae
Phascolarctobacterium
Parabacteroides
Lactobacillus
Anaerostipes
Streptococcus
Clostridiales
Clostridium
Enterococcus
Collinsella
Megamonas
Subdoligranulum
Paraprevotella
Coprobacillus
Anaerotruncus
Klebsiella
Sporacetigenium
Holdemania
Mitsuokella
Sutterella
Acidaminococcus
Bilophila
Veillonella
Methanosphaera
Parasutterella
Burkholderiales
Peptostreptococcaceae
Desulfovibrio
Leuconostoc
Enterobacter
Marvinbryantia
Turicibacter
Eggerthella
Fusobacterium
Citrobacter
Gordonibacter
Porphyromonas
Salmonella
Pediococcus
Actinomyces
Enterobacteriaceae
Parvimonas
Peptostreptococcus
Weissella
Finegoldia
−3
−2
−1
0
1
2
Relative abundance ($\log_{10}$)

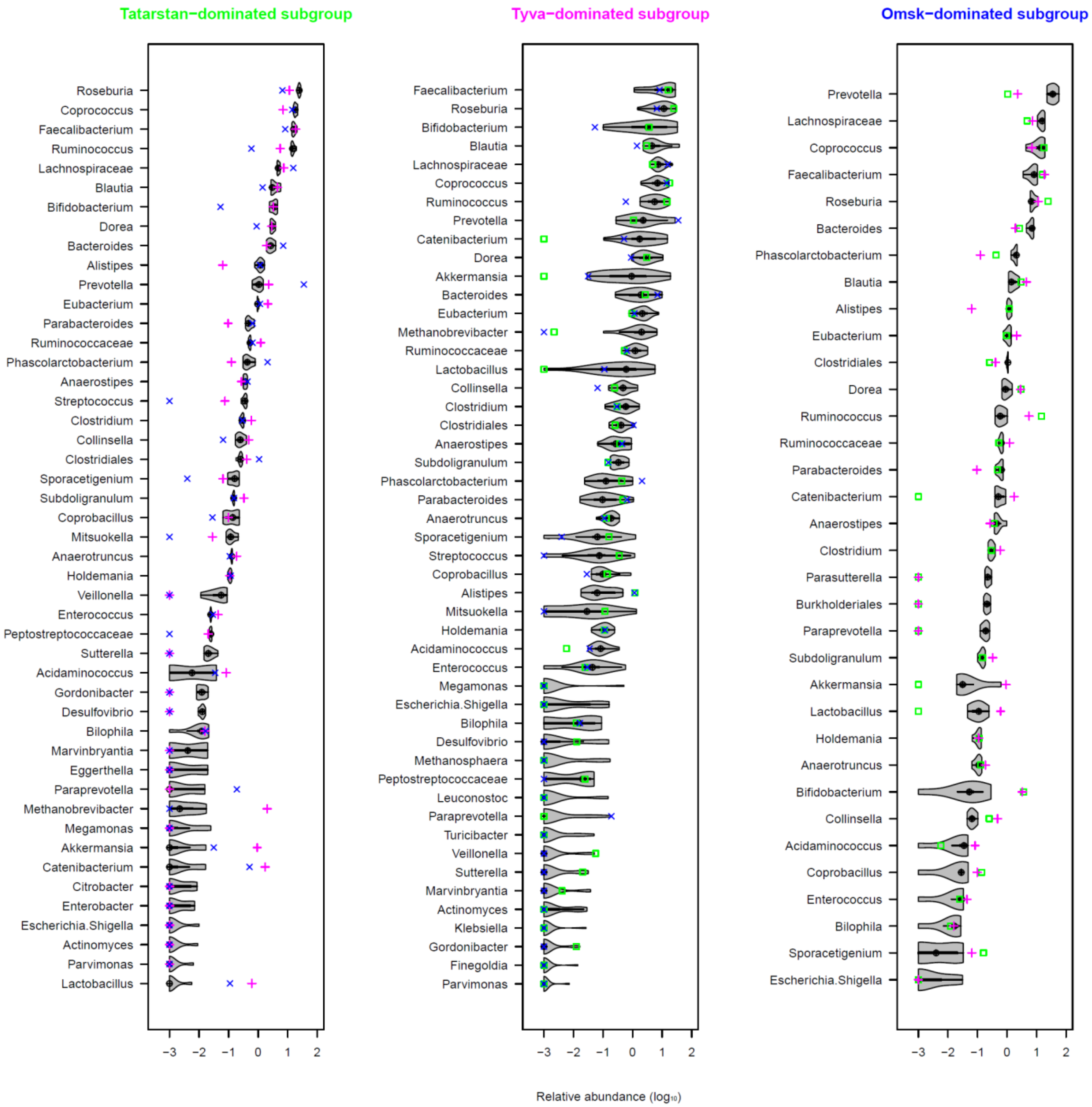
b.



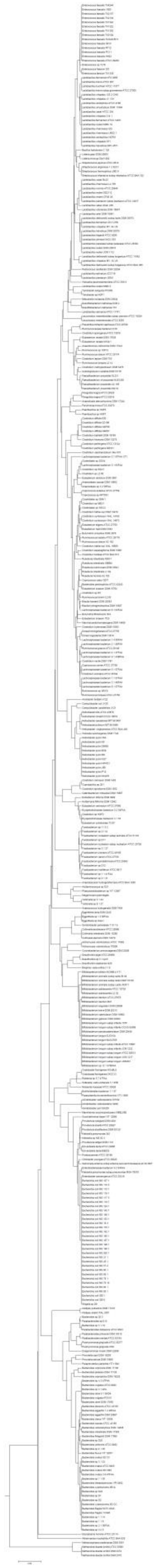
Supplementary Fig. S4 – The microbial composition of 96 Russian samples. a, Violin plots of the relative abundance ($\log_{10}$) of 61 microbial genera with non-zero coverage. Violin plots are an extended version of boxplots for the visualisation of probability density. Circles denote median. In order to avoid calculating logarithm of zero abundances, pseudocounts of 0.001% to abundance matrices were added. Thus, value -3 here means absence of the genus **b**, Hierarchical heat plot constructed using a Spearman correlation-based dissimilarity metric and Ward linkage. The colour bar on the side denotes the type of settlement for all samples.



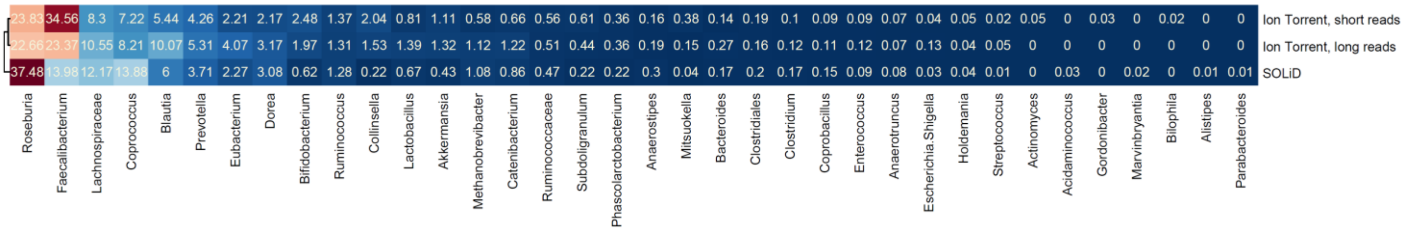
Supplementary Fig. S5 - P-value clustering of genus abundance reveals several significant tight subgroups, with the three largest being associated with rural locations in the Tyva (TYV), Tatarstan (TAT) and Omsk regions (OM). Clustering was performed with the *pvclust* R package with a Spearman correlation-based metric and Ward linkage and represents 1,000 bootstrap repetitions. Red boxes highlight clusters with an AU (approximately unbiased) p-value > 98.5%. At each node of the tree, the AU, bootstrap probability and edge number values are indicated in red, green and grey, respectively.



Supplementary Fig. S6- Bacterial composition of the three largest compact subgroups, as revealed by *pvclust*. These subgroups are composed of samples from Tatarstan (n = 8), Tyva (n = 15) and Omsk (n = 7) correspondingly. For each subgroup, the genera showing non-zero abundance are arranged in decreasing order of the means. Violin plots of relative abundance ($\log_{10}$ of percentage) are an extended version of boxplots for the visualisation of probability density. A pseudocount of 0.001% to original data set was added before the analysis in order to avoid calculating logarithm of zero abundances, thus -3 value here means absence of genus. Each of the violin plots contains the median values for the corresponding genus in the other two subgroups (green square-Tatarstan, magenta cross-Tyva, blue plus-Omsk).



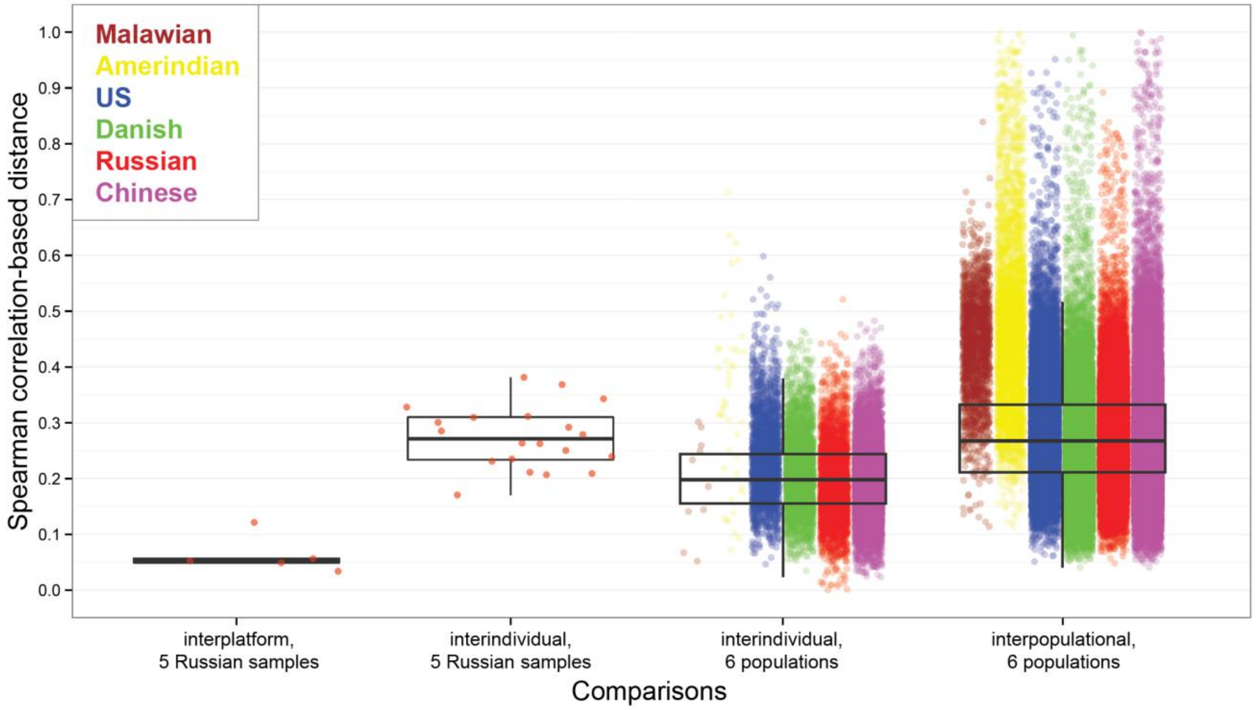
Supplementary Fig. S7 - 16S rRNA phylogenetic tree of reference genomes used in UniFrac calculations (constructed in MUSCLE).



Roseburia	Faecalibacterium	Lachnospiraceae	Coprococcus	Blautia	Prevotella	Eubacterium	Dorea	Bifidobacterium	Ruminococcus	Collinsella	Lactobacillus	Akkermansia	Methanobrevibacter	Catenibacterium	Ruminococcaceae	Subdoligranulum	Phascolarctobacterium	
23.83	34.56	8.3	7.22	5.44	4.26	2.21	2.17	2.48	1.37	2.04	0.81	1.11	0.58	0.66	0.56	0.61	0.36	Ion Torrent, short reads
22.66	23.37	10.55	8.21	10.07	5.31	4.07	3.17	1.97	1.31	1.53	1.39	1.32	1.12	1.22	0.51	0.44	0.36	Ion Torrent, long reads
37.48	13.98	12.17	13.88	6	3.71	2.27	3.08	0.62	1.28	0.22	0.67	0.43	1.08	0.86	0.47	0.22	0.22	SOLiD

Anaerostipes	Mitsuokella	Bacteroides	Clostridiales	Clostridium	Coprobacillus	Enterococcus	Anaerotruncus	Escherichia.Shigella	Holdemania	Streptococcus	Actinomyces	Acidaminococcus	Gordonibacter	Marvinbryantia	Bilophila	Alistipes	Parabacteroides	
0.16	0.38	0.14	0.19	0.1	0.09	0.09	0.07	0.04	0.05	0.02	0.05	0	0.03	0	0.02	0	0	Ion Torrent, short reads
0.19	0.15	0.27	0.16	0.12	0.11	0.12	0.07	0.13	0.04	0.05	0	0	0	0	0	0	0	Ion Torrent, long reads
0.3	0.04	0.17	0.2	0.17	0.15	0.09	0.08	0.03	0.04	0.01	0	0.03	0	0.02	0	0.01	0.01	SOLiD

Supplementary Fig. S8 - Comparison of sequencing technologies and different read lengths: abundance plot of prevalent microbial genera. For one sample (ID: X_77), whole-genome shotgun sequencing was performed using the SOLiD and Ion Torrent platforms with two different read lengths (short 120±15 and long 229±58 bp). Values are given as percentages of total relative abundance.



Supplementary Fig. S9– Comparison of the size of the effect of sequencing platform choice on the microbial composition with the size of the effect of difference between the individuals and the difference between the populations. The four boxplots combined with the scatterplots show the distributions of Spearman correlation-based distances for the following groups (the colours denote the nation): **a,** the interplatform distances between SOLiD and Illumina Russian metagenomes for the *same* sample, for 5 pairs of samples: 0.05±0.03 (median±s.d.); **b,** the interindividual distances between SOLiD and Illumina Russian metagenomes for *different* samples, for 5 samples: 0.26±0.10; **c,** the interindividual distances in each population (the distances between all possible pairs of samples within the population): the distance across all populations is 0.20±0.07; **d,** the interpopulational distances (the distances between all possible pairs of samples, where one sample is located in the fixed population and the other is in any of the other populations): the distance across all populations is 0.28±0.11. The paired interplatform distances for 5 samples (**a**) are significantly lower than the respective interindividual distances (**b**) (one-sided Mann-Whitney test, P = 0.00085), the interindividual distances in each population (**c**) ($P \leq 0.008$) and the interpopulational distances in each population (**d**) ($P \leq 0.00014$).

Supplementary Tables

Supplementary Table S1 – Analysis of similarities (ANOSIM) between samples from Russia and other countries (10,000 permutations).

Comparison	R statistic	P-value
Russian vs. US	0.74	9.999×10^{-5}
Russian vs. Chinese	0.5	9.999×10^{-5}
Russian vs. Danish	0.26	9.999×10^{-5}
Russian vs. Malawian	0.041	0.37
Russian vs. Amerindian	-0.038	0.67

Supplementary Table S2 – The Russian cohort exhibits the lowest fraction of the microbiota composition driven by *Prevotella* or *Bacteroides*.

	Fraction of samples having Prevotella or Bacteroides as genus #1 by abundance, %	**Fraction of samples having >35% of Prevotella or Bacteroides, %**
Russia	**37.5**	**18.8**
Denmark	83.5	61.1
China	95.7	90.0
USA	96.4	88.3
Malawi	80.0	80.0
Venezuela	70.0	40.0

Supplementary Table S3 - Average silhouette width for PAM clustering of the Russian and global samples using genera relative abundance and various dissimilarity metrics.

Metric	ASW (genera, Russian)	ASW (genera, Russian and non-Russian)
JSD	0.187	0.311
Spearman correlation-based	0.156	0.237
Euclidean	0.308	0.421
Manhattan	0.190	0.384
Canberra	0.086	0.142
Bray-Curtis	0.194	0.387
UniFrac	0.349	0.414

Supplementary Table S4 – Comparison of age and BMI between the Russian and non-Russian cohorts. Age can significantly influence the microbiota composition; in particular, Bacteroidetes can dominate the microbiota of elderly people compared to younger humans[8]. However, the prevalence of Bacteroidetes in the US, Danish and Chinese groups in comparison with the Russian samples is not primarily due to an age effect because the age (as well as BMI values) ranges of the groups were comparable. Figures for the US cohort were taken from the Human Microbiome Project enrolment criteria.

	Age, years			**BMI**		
	min	**max**	**mean**	**min**	**max**	**mean**
African	23	27	25.1	20	24.2	21.5
Amerindian	5	53	20.1	14.1	26.1	20.8
Chinese	21	70	48	15.6	31.4	23.4
Danish	42	69	57	18.6	40.2	27.8
Russian	14	85	40	16	36.1	23.9
US	18	40	N/A	18	35	N/A

Supplementary Note 1

Comparison of taxonomic profiling with approach based on clade-specific gene detection

To validate techniques for assessing taxonomic composition, colour-space reads of 96 Russian samples were converted to basespace and classified using MetaPhlAn (Supplementary Data 8). The relative abundances of 58 genera found in both of these genera sets were highly correlated between the methods (Spearman correlation 0.86±0.04). Presumably, disconcordance between the methods was mostly attributed to different treatments of few taxonomically ambiguous genomes: for example, MetaPhlAn puts *Bacteroides pectinophilus* into the Bacteroidaceae family and *Eubacterium rectale* into the *Eubacterium* genus, while our method classified these genomes as belonging to the Lachnospiraceae family and *Roseburia* genus, respectively (basing on 16S rRNA classification).

Supplementary Note 2

Comparison of taxonomic profiling across different sequencing technologies (SOLiD, Ion Torrent, 454 and Illumina)

The similarity of metagenomes resulting from whole-genome sequencing on different platforms was confirmed through a series of experiments.

a) One of the Russian samples (Spb_61_1P) sequenced on the SOLiD 4 platform was also processed using the Ion Torrent and 454 GS FLX+ platforms. The resulting relative genus abundance vectors were highly correlated pairwise (Spearman correlation for SOLiD vs. 454 - 0.94, Ion Torrent vs. 454 - 0.93, SOLiD vs. Ion Torrent - 0.99; mean correlation across the Russian cohort obtained using SOLiD - 0.78±0.04 s.d.). Moreover, variation of read length on Ion Torrent platform resulted in a composition similar to that obtained via the SOLiD platform (two different read length ranges (120±15 bp and 229±58 bp) for Russian sample X77: Spearman correlation with SOLiD - 0.95±0.02 s.d.; see Supplementary Fig. S9).

b) Five of the Russian samples (Kh_249, NOV_284, Spb_66_6P, TAT_130 and TYV_212) sequenced on the SOLiD platform were additionally processed on the Illumina HiSeq 2000 platform. The samples were selected such that the different poles of taxonomic diversity discovered in the Russian cohort were represented, i.e., samples dominated by *Prevotella*, *Bacteroides*, Firmicutes and Actinobacteria. The resulting high correlation of the genus compositions obtained using the SOLiD and Illumina platforms (Spearman 0.93±0.03) demonstrated their similarity.

The size of the effect of sequencing platform choice on the microbial composition was lower than the size of the effect of difference between the individuals and the difference between the populations (Supplementary Fig. S8, S9).

Supplementary Note 3

Microbial community of Russian metagenomes

For Russian samples, the fraction of reads that aligned to reference sets was comparable to values obtained for non-Russian datasets (Supplementary Table S2). Of 86 microbial genera in the reference set, 61 were present in at least one sample. The main quantitative dominants were the genera *Prevotella*, *Roseburia*, *Bacteroides*, *Faecalibacterium*, unclassified Lachnospiraceae, *Coprococcus*, *Blautia* and *Ruminococcus*, together comprising >80% of summary relative abundance across all Russian samples. The validity of our genome catalogue for taxa abundance assessment was supported by the high correlation of the associated genera proportions for the Danish samples with the values obtained for these metagenomic samples in the MetaHIT study on a different genome set and read alignment software[13] (Spearman correlation 0.83±0.03).