Environmental biodiversity, human microbiota and allergy are interrelated

Ilkka Hanski, Leena von Hertzen, Nanna Fyhrquist, Kaisa Koskinen, Kaisa Torppa, Tiina Laatikainen, Piia Karisola, Petri Auvinen, Lars Paulin, Mika J. Mäkelä, Erkki Vartiainen, Timo U. Kosunen, Harri Alenius and Tari Haahtela

Taking into account variation in sample size while analyzing association between bacterial diversity and atopy

In the main text, we use the total number of all bacterial genera in the samples as a covariate while analyzing associations between atopy and generic diversity in the six main bacterial classes (Table 1 in the main text). The total number of bacterial genera in the sample is interpreted as a biological measure of sample size. Commonly, rarefaction is used to correct for variation in sample size, that is, a constant number of units (here DNA sequences) is randomly drawn from each original sample to make them comparable. However, this approach is problematic in the present context due to the highly uneven distribution of relative abundances of the bacterial genera, which is partly caused by technical reasons. For instance, a specific genus may be very common in a particular sample because of bias in the PCR reaction. Correcting for sample size by rarefaction may hence lead to biased estimates of generic diversity. In our case, when we rarefied the samples to the constant size of 4567 sequences per sample, the only significant difference at the 5% level in generic diversity between healthy and atopic individuals was in Actinobacteria ($P = 0.038$). No significant differences were obtained for the other bacterial classes, including Gammaproteobacteria ($P = 0.079$). For Actinobacteria the result remained the same as when variation in sample sizes was accounted for by using the total number of bacterial genera as a covariate (see Table 1 in the main text). For Gammaproteobacteria the result changed greatly, and a simulation study was conducted to demonstrate that correcting for variation in sample size via rarefaction dramatically reduces statistical power to detect a true difference for a rare class, such as Gammaproteobacteria in the present study, which comprises only 3% of all sequences in the data.

The following simulation study was conducted to demonstrate the bias arising from rarefaction. We assumed a pool of 1000 genera from which samples were drawn for 100 individuals. Out of the 1000 genera, the first 50 ones were designated to belong to class G, which corresponds to Gammaproteobacteria in the empirical samples. The abundance distribution of the 1000 genera in the pool was defined by a truncated lognormal variate Y~exp(X), where X is normally distributed with mean 1 and SD 5, truncated at the value 1. We drew with replacement 100 samples from the pool, where the size of each sample was a random variable uniformly distributed between 5000 and 25000 sequences as in the empirical data. For half of the samples out of the total 100 (corresponding to 50 individuals), we doubled the number of sequences in the class G genera, to mimic higher relative abundance of these taxa in “healthy” (H) individuals. The remaining 50 samples represented “atopic” (A) individuals. We thereafter analyzed the data by regression to explain the number of genera in class G by the two covariates, the total number of genera per sample and the type of the sample, H or A (analogously to Fig. 2b in the main text). To demonstrate the biasing effect due to rarefaction, a random subsample of 5000 sequences was drawn from each of the 100 generated samples, whereafter anova was used to test for the difference in the number of genera within class G between H and A individuals. Figure S1 shows a representative example of the simulation results. The effect of individual type (H or A) on the number of genera in class G is highly significant in the regression model (t = 3.50, P = 0.0007). In contrast, the difference in the number of genera in class G between H and A individuals was not significant after rarefaction correction (P = 0.051). These results are very similar to the empirical results (Fig. 2b in the main text and above). If a greater difference in the relative abundances (number of sequences) of class G genera between H and A individuals is assumed, even rarefaction would detect the difference, but the example in Fig. S1 shows that accounting for variation in sample size by using the total number of bacterial genera as a covariate allows detection of smaller true differences, i.e. increases the statistical power of the analysis.

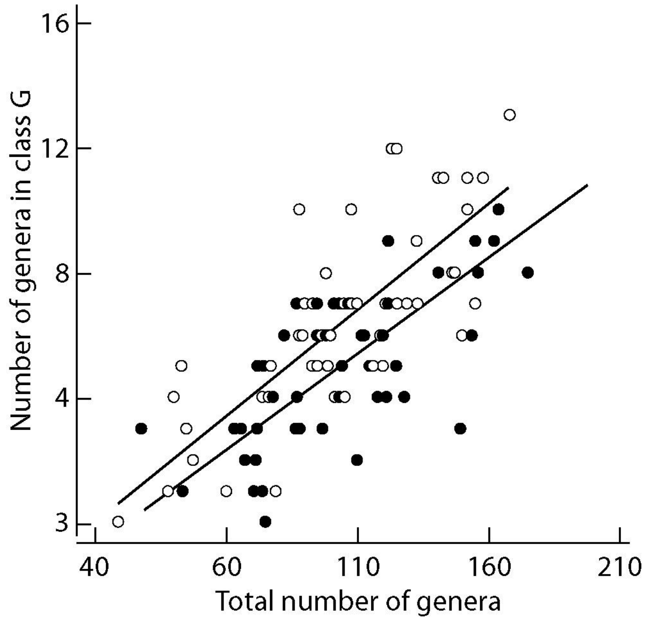

Fig. S1. An example based on simulated data, demonstrating the use of the total number of bacterial genera as a covariate to test for a difference in the number of genera within a particular class of bacteria. The test compares two types of individuals marked by filled (A) and open circles (H). See the text above for further details and compare with Fig. 2b in the main text, which gives the empirical result.

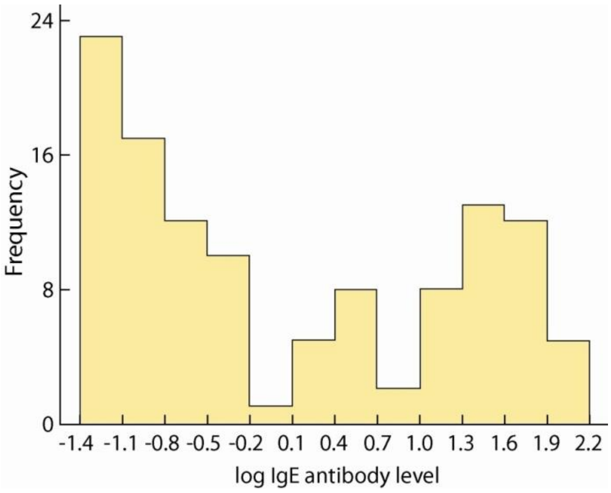

Fig. S2. Distribution of IgE antibody levels in the study population. The distribution is bimodal on a logarithmic scale. We identified the two modes of the distribution as “healthy” versus “atopic” individuals using the cut-off level of 2.5 kU_A/l (corresponding to 0.4 on the logarithmic scale). Analyses were repeated with the alternative cut-off point of 1.0 (0.0 on the logarithmic scale).

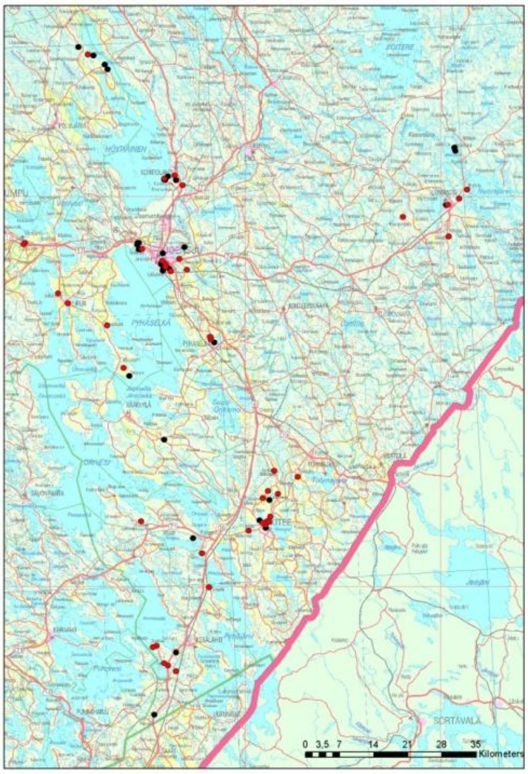

Fig. S3. Map of the study area in eastern Finland. The map shows the locations of the homes of atopic (black dot) and healthy individuals (open symbols).

Table S1. Principal component analysis of the five land use types. *n* = 95, including study subjects for which there are data for both the skin microbiota and the land use types.

Factor	Vectors			
	1	2	3	4
Eigenvalue	1.94	1.12	1.11	0.77
% of variance	38.9	23.7	22.1	15.4
Correlations				
Agricultural land	0.388	0.061	-0.637	0.577
Forest	0.626	0.019	0.276	-0.446
Built areas	-0.383	0.668	0.332	0.292
Lakes, water bodies	-0.558	-0.404	-0.352	-0.292
Wetlands	0.009	-0.622	0.533	0.546

Table S2. Principal component analysis of the numbers of genera in the six main bacterial classes. *n* = 95, including study subjects for which there are data for both the skin microbiota and the land use types.

Factor	Vectors					
	1	2	3	4	5	6
Eigenvalue	3.35	1.00	0.66	0.46	0.29	0.25
% of variance	55.8	16.7	11.0	7.6	4.8	4.1
Correlations						
Actinobacteria	-0.432	-0.171	0.589	0.104	0.428	-0.494
Bacilli	-0.413	-0.389	-0.086	0.709	-0.254	0.323
Clostridia	-0.313	-0.673	-0.270	-0.610	0.004	0.061
Betaproteobacteria	-0.439	0.324	-0.411	-0.012	-0.428	-0.592
Alphaproteobacteria	-0.415	0.333	0.506	-0.338	-0.399	0.433
Gammaproteobacteria	-0.425	0.389	-0.385	-0.003	0.641	0.332

Table S3. Statistics for the five groups of plants recorded in 114 yards. The plant species have been divided into common species (forming one or more distinct patches of vegetation) versus uncommon species (distributed sparsely as single individuals). Pteridophytes had a small number of species and they were not analyzed further. The last two columns give the regression coefficient and the *P* value for the effect of atopy on the number of plant species in the particular category, using the total number of plant species in the yard as a covariate (as in Fig. 2a for the uncommon native flowering plants). Atopic individuals were scored as 1 and healthy individuals as 0.

Plant group	Category	Number of species					
		mean	sd	min	max	coeff	*P*
Trees and shrubs	common	5.25	4.08	0	23	-0.15	0.82
	uncommon	14.8	6.96	2	33	0.64	0.50
Pteridophytes	common	0.58	1.09	0	5		
	uncommon	1.60	1.23	0	6		
Grasses and sedges	common	6.82	3.50	0	17	-0.51	0.43
	uncommon	4.13	2.38	0	13	0.02	0.96
Flowering plants	common	24.5	13.6	2	62	-0.01	0.99
	uncommon	31.7	10.8	12	58	-5.30	0.0022
Decorative plants	common	5.14	6.44	0	26	0.32	0.73
	uncommon	24.1	16.8	0	82	3.56	0.10

Table S4. Characteristics of the study subjects and their living conditions in 2003. Individuals have been divided into atopic and healthy ones based on skin prick testing performed in 2003 (*n* = 112). The table gives the numbers of individuals and percentages (in brackets) in the two groups. The effects of the type (4 categories), age (4) and condition (3) of the house were determined in 2010, and in these cases atopy was determined by the IgE screen in 2010. The *P* value is for chi-squared test of independence.

	Healthy	Atopic	*P*
Mean age (with sd)	8.9 (1.6)	8.5 (1.4)	0.15
Sex ratio (females)	39 (60.0)	27 (57.5)	0.79
Type of house			0.72
Age of house			0.38
Condition of house			0.66
Indoor exposure to tobacco smoke	37 (56.9)	21 (44.7)	0.20
Parental farming in the past year	14 (21.5)	6(12.8)	0.23
Parental farming, current	10 (15.4)	7(14.9)	0.94
Indoor pets within 10 years	30 (46.2)	23 (48.9)	0.77
Current contacts with domestic animals			
Cow (*n*=111)	24 (36.9)	19 (41.3)	0.64
Horse (*n*=111)	27 (41.5)	20 (43.5)	0.84
Dog	60 (92.3)	43 (91.5)	0.88
Cat	56 (86.2)	41 (87.2)	0.87
Visits to a stable in the past year	31 (47.7)	21 (44.7)	0.75
Physician-diagnosed atopic disease			
Asthma	1 (1.5)	6 (12.8)	0.02
Hay fever (*n*=110)	2 (3.1)	4 (8.9)	0.19
Atopic eczema	13 (20.0)	18 (38.3)	0.03
Parental history of atopic disease			
Atopy (SPT), mother	18(29.0)	18(39.0)	0.27
Asthma (self-reported)			
mother (*n*=111)	4 (6.3)	8 (17.0)	0.07
father (*n*=102)	3 (5.0)	1 (2.4)	0.50
Hay fever (self-reported)			
mother (*n*=111)	4 (6.3)	4 (8.5)	0.65
father (*n*=103)	7 (11.7)	1 (2.3)	0.08
Atopic eczema (self-reported)			
mother (*n*=111)	5 (7.8)	15 (31.9)	0.0011
father (*n*=101)	4 (6.7)	3 (7.3)	0.90

Table S5. Specific IgE tests against common inhalant allergens. This table gives the number and percentage of study subjects out of 118 who had a positive test result ($\geq$0.35 kU$_A$/l) against the specific inhalent allergen. The next two columns give the median and maximum test result for positive individuals, and the last two columns give the effect of atopy as defined by the specific allergen on the generic diversity of gammaproterobacteria on the subject's skin (as in Fig. 2b for atopy defined by the generic Phadiatop© screen for a mixture of common inhalant allergens).

Allergen	Positive cases	Percentage positive	median of positive	maximum of positive	coeff	*P*
Cat	31	26%	2.95	99.0	-0.79	0.094
Dog	37	31%	1.65	44.0	-1.05	0.017
Horse	14	12%	2.61	10.1	-1.33	0.042
Birch	33	28%	31.3	596.0	-0.89	0.052
Timothy grass	38	32%	13.4	717.0	-0.97	0.027
Mugwort	24	20%	1.77	8.6	-1.07	0.035

Table S6. Logistic regression models of atopy for three different definitions of atopy: Model 1, atopy defined using the IgE antibody level >2.5 kU$_A$/l (see Fig. S2); Model 2, as Model 1 but with the IgE cut-off value 1 kU$_A$/l; and Model 3, atopy defined based on skin prick testing (SPT) conducted in 2003. The columns give the coefficients of the logistic model and their *P* values.

Variable	Model 1		Model 2		Model 3	
	coeff	*P*	coeff	*P*	coeff	*P*
Constant	-0.58	0.023	-0.36	0.13	-0.55	0.023
Land use types, PC1$_{env}$	-0.52	0.0059	-0.49	0.0059	-0.31	0.086
Flowering plants (res)	-0.10	0.0016	-0.08	0.0076	-0.08	0.0069
Gammaproteobacteria	-0.31	0.015	-0.21	0.082	-0.27	0.027
P value of the model	0.20		0.081		0.085	
positive cases/*N*	38/94		41/94		36/91	

Table S7. Associations between the relative abundance and generic diversity of the six main bacterial classes and IL-10 expression separately in healthy (*n* = 45) and atopic individuals (*n* = 25). Relative abundance and IL-10 expression were log transformed. The table gives the *P* values from linear regression as in Table 1 in the main text. In the case of generic diversity, the total number of bacterial genera was used as a covariate (as in Fig. 2b in the main text).

	relative abundance		generic diversity	
bacterial class	healthy	atopic	healthy	atopic
Actinobacteria	0.662	0.264	-0.887	0.148
Bacilli	-0.830	-0.932	-0.685	0.799
Clostridi	0.850	-0.928	0.858	-0.355
Betaproteobacteria	0.236	-0.209	-0.904	-0.316
Alphaproteobacteria	0.573	-0.285	-0.818	-0.713
Gammaproteobacteria	0.015	-0.304	0.529	-0.145

Table S8. Associations between the relative abundance of the 12 most common genera of gammaproteobacteria (average relative abundance > 0.001%) and IL-10 expression separately in healthy (*n* = 45) and atopic individuals (*n* = 25). Both variables were log transformed. The last two columns table give the *P* values from linear regression as in Table 1 in the main text.

Genus	average relative abundance	healthy	atopic
Acinotebacter	0.527	0.0004	-0.139
Enhydrobacter	1.259	0.039	0.551
Moraxella	0.031	0.531	0.790
Pseudomonas	0.146	0.217	0.704
Pantoea	0.015	-0.871	-0.291
Aggregatibacter	0.018	-0.106	-0.842
Haemophilus	0.067	0.570	-0.861
Luteimonas	0.037	0.228	-0.785
Rhodanobacter	0.030	-0.114	-0.318
Lysobacter	0.012	0.928	-0.158
Dyella	0.053	-0.549	-0.279
Stenotrophomonas	0.127	0.070	0.788