Oikos

OIK-07280

Ellis, V. A., Huang, X., Westerdahl, H., Jönsson, J., Hasselquist, D., Neto, J. M., Nilsson, J.-Å., Nilsson, J., Hegemann, A., Hellgren, O. and Bensch, S. 2020. Explaining prevalence, diversity, and host specificity in a community of avian haemosporidian parasites. – Oikos doi: 10.1111/oik.07280

Appendix 1

Text A1. We split the datasets up (see descriptions of Datasets A-D in the main text) in order to strike a balance between controlling for error associated with small sample sizes and maintaining statistical power in our tests. The minimum of 16 individual birds (Datasets C and D) aims for the 10-20 individuals that Jovani and Tella (2006) suggested were adequate for minimizing errors associated with estimating prevalence. In fact, this was not just a statistical decision but a sampling decision – we aimed to sample species to this threshold. As described in the main text, the species with 16 or more sampled individuals represent 97% of our total samples. Thus, tests of hypotheses H1–H5 which used Dataset C are unlikely to change substantially with changes in the minimum sample size used. Nevertheless, we re-ran H1–H5 using minimum sample sizes of 5 and 25 to see how the results would be affected.

Hypothesis H1 states that the prevalence and diversity of parasites are positively related to host abundance. We found no evidence of this in our analysis and that result held when using a sample size of 5 (prevalence, *Plasmodium* $\beta = 0.31$, $p = 0.15$, *Haemoproteus* $\beta = 0.05$, $p = 0.79$, *Leucocytozoon* $\beta = -0.03$, $p = 0.85$; diversity, *Plasmodium* $\beta = 0.05$, $p = 0.19$, *Haemoproteus* $\beta = 0.02$, $p = 0.61$, *Leucocytozoon* $\beta = -0.04$, $p = 0.21$) and 25 (prevalence, *Plasmodium* $\beta = 0.28$, $p = 0.36$, *Haemoproteus* $\beta = -0.07$, $p = 0.70$, *Leucocytozoon* $\beta = -0.09$, $p = 0.69$; diversity, *Plasmodium* $\beta = 0.03$, $p = 0.67$, *Haemoproteus* $\beta = -0.03$, $p = 0.57$, *Leucocytozoon* $\beta = -0.04$, $p = 0.34$).

Hypothesis H2 states that the prevalence and diversity of parasites are positively related to host body size. We found no evidence for this in our analysis and that result held when using a sample size of 5 (prevalence, *Plasmodium* $\beta = 0.43$, $p = 0.71$, *Haemoproteus* $\beta = 0.49$, $p = 0.72$, *Leucocytozoon* $\beta = 0.52$, $p = 0.70$; diversity, *Plasmodium* $\beta = 0.10$, $p = 0.46$, *Haemoproteus* $\beta = 0.05$, $p = 0.72$, *Leucocytozoon* $\beta = 0.16$, $p = 0.19$) and 25, although the p-value for diversity of *Leucocytozoon* went from 0.17 (original analysis) to 0.06 (prevalence, *Plasmodium* $\beta = 0.98$, $p = 0.44$, *Haemoproteus* $\beta = -0.10$, $p = 0.89$, *Leucocytozoon* $\beta = 0.34$, $p = 0.80$; diversity, *Plasmodium* $\beta = 0.16$, $p = 0.39$, *Haemoproteus* $\beta = 0.10$, $p = 0.55$, *Leucocytozoon* $\beta = 0.26$, $p = 0.06$).

Hypothesis H3 states that the prevalence and diversity of parasites are related to host habitat preferences, here defined by two variables (PC1 corresponding to habitat structure and PC2 corresponding to habitat wetness; see main text). We found a relationship between *Haemoproteus* prevalence and host habitat structure and *Haemoproteus* diversity and host habitat wetness; *Leucocytozoon* prevalence was related to habitat structure. Using a sample size of 5, the P value for *Leucocytozoon* prevalence in relation to host habitat structure went up to 0.08 from 0.05 and the P value for the relationship between *Haemoproteus* diversity and host habitat wetness went to 0.11 from 0.04 (prevalence, *Plasmodium*, PC1 $\beta = -0.03$, $p = 0.87$, PC2 $\beta = 0.28$, $p = 0.19$, *Haemoproteus* PC1 $\beta = -0.36$, $p = 0.02$, PC2 $\beta = 0.33$, $p = 0.06$, *Leucocytozoon* PC1 $\beta = -0.41$, $p = 0.08$, PC2 $\beta = 0.25$, $p = 0.34$; diversity, *Plasmodium* PC1 $\beta = -0.02$, $p = 0.45$, PC2 $\beta = 0.01$, $p = 0.79$, *Haemoproteus* PC1 $\beta = -0.02$, $p = 0.27$, PC2 $\beta = -0.04$, $p = 0.11$, *Leucocytozoon* PC1 $\beta = -0.03$, $p = 0.11$, PC2 $\beta = 0.01$, $p = 0.50$). Using a sample size of 25, *Haemoproteus* prevalence was also related to habitat wetness (the p-value went from 0.09 in the original analysis to 0.02) and the p-value for *Haemoproteus* diversity in relation to habitat wetness went from 0.04 to 0.14 (prevalence, *Plasmodium*, PC1 $\beta = -0.16$, $p = 0.49$, PC2 $\beta = 0.12$, $p = 0.61$, *Haemoproteus* PC1 $\beta = -0.35$, $p = 0.04$, PC2 $\beta = 0.46$, $p = 0.02$, *Leucocytozoon* PC1 $\beta = -0.34$, $p = 0.16$, PC2 $\beta = 0.30$, $p = 0.29$; diversity, *Plasmodium* PC1 $\beta = -0.01$, $p = 0.66$, PC2 $\beta = 0.00$, $p = 0.93$, *Haemoproteus* PC1 $\beta = -0.03$, $p = 0.31$, PC2 $\beta = -0.05$, $p = 0.14$, *Leucocytozoon* PC1 $\beta = -0.03$, $p = 0.23$, PC2 $\beta = 0.01$, $p = 0.57$).

Hypothesis H4 states that the prevalence and diversity of parasites are related to the evolutionary distinctiveness of hosts. We found negative relationships between *Haemoproteus* prevalence and diversity and host evolutionary distinctiveness. Our results held using a sample size of 5, although the p-value for *Haemoproteus* diversity went up to 0.11 from 0.03 while the estimate of the coefficient remained the same (prevalence, *Plasmodium* $\beta = 0.00$, $p = 0.98$, *Haemoproteus* $\beta = -0.12$, $p = 0.02$, *Leucocytozoon* $\beta = 0.01$, $p = 0.80$; diversity, *Plasmodium* $\beta = 0.00$, $p = 0.69$, *Haemoproteus* $\beta = -0.01$, $p = 0.11$, *Leucocytozoon* $\beta = 0.00$, $p = 0.92$). We had similar results for a sample size of 25 with the p-value of Haemoproteus diversity going up to 0.09 (prevalence, *Plasmodium* $\beta = 0.05$, $p = 0.49$, *Haemoproteus* $\beta = -0.18$, $p = 0.002$, *Leucocytozoon* $\beta = 0.07$, $p = 0.33$; diversity, *Plasmodium* $\beta = 0.01$, $p = 0.53$, *Haemoproteus* $\beta = -0.01$, $p = 0.09$, *Leucocytozoon* $\beta = 0.01$, $p = 0.27$).

Hypothesis H5 states that host phylogenetic relationships are related to parasite prevalence, diversity, and lineage composition. To test this we calculated phylogenetic signal (using Pagel's λ and Blomberg's K) for the prevalence and diversity of parasites among host species. We found no signal for *Plasmodium* prevalence and diversity, but *Haemoproteus* (Blomberg's K) and *Leucocytozoon* (Pagel's λ and Blomberg's K) prevalence had significant phylogenetic signal;

neither *Haemoproteus* nor *Leucocytozoon* diversity showed signal. We also tested the correlation between the host phylogenetic distance matrix and the Bray–Curtis dissimilarity of parasite lineage composition among hosts using Mantel tests. We found significant correlations between host phylogenetic distance and parasite lineage composition for *Haemoproteus* and *Leucocytozoon*, but not *Plasmodium* ($p = 0.08$). Using a sample size of 5, the p-value for Pagel's λ of *Plasmodium* prevalence dropped to 0.02 from 0.12; the other results were consistent with the results of our original analysis (*Plasmodium*, prevalence, $\lambda = 0.66$, $p = 0.02$, $K = 0.49$, $p = 0.10$, diversity, $\lambda = 0.27$, $p > 0.99$, $K = 0.40$, $p = 0.21$, composition, $r = 0.09$, $p = 0.08$; *Haemoproteus*, prevalence, $\lambda = 0.31$, $p > 0.99$, $K = 0.58$, $p = 0.01$, diversity, $\lambda < 0.01$, $p > 0.99$, $K = 0.44$, $p = 0.11$, composition, $r = 0.26$, $p < 0.01$; *Leucocytozoon*, prevalence, $\lambda > 0.99$, $p < 0.01$, $K = 0.59$, $p = 0.09$, diversity, $\lambda = 0.40$, $p = 0.21$, $K = 0.41$, $p = 0.21$, composition, $r = 0.21$, $p = 0.03$). Using a sample size of 25, the p-value for the Mantel test of *Plasmodium* lineage composition dropped to 0.04 from 0.08 (in the original analysis) and the p-value for Blomberg's K for *Haemoproteus* prevalence increased to 0.07 from 0.03 (in the original analysis; *Plasmodium*, prevalence, $\lambda = 0.52$, $p = 0.09$, $K = 0.68$, $p = 0.11$, diversity, $\lambda = 0.09$, $p = 0.82$, $K = 0.53$, $p = 0.27$, composition, $r = 0.13$, $p = 0.04$; *Haemoproteus*, prevalence, $\lambda < 0.01$, $p > 0.99$, $K = 0.63$, $p = 0.07$, diversity, $\lambda < 0.01$, $p > 0.99$, $K = 0.48$, $p = 0.48$, composition, $r = 0.30$, $p < 0.01$; *Leucocytozoon*, prevalence, $\lambda > 0.99$, $p < 0.01$, $K = 1.96$, $p < 0.01$, diversity, $\lambda = 0.51$, $p = 0.08$, $K = 0.65$, $p = 0.09$, composition, $r = 0.19$, $p = 0.04$).

Hypothesis P1 uses Dataset D (host species at minimum sample size of 16 individuals and parasites with a minimum of 10 records within those hosts) and is a test of the phylogenetic relationships of the host species of each parasite (a separate test was conducted for each parasite lineage; all but two *Plasmodium* lineages infected more closely related host species than expected by chance). Lowering the minimum parasite records will introduce more parasites into the analysis (if minimum parasite records are reduced to 5, 10 new lineages will be added to the analysis), but the fact that all but two *Plasmodium* parasites infected more closely related hosts than expected by chance suggests that the main result (most parasites infect closely related hosts) is unlikely to change. Nevertheless, we report that when the minimum number of parasite records is reduced to 5, the 10 additional lineages all have more closely related hosts than expected by chance ($p < 0.01$ for all). Increasing the minimum parasite records will only remove parasites from the analysis and that cannot change the results (e.g. the most abundant parasites in the community are the three *Haemoproteus majoris* lineages which all infect more closely related host species than expected by chance). The test of hypothesis P2 also used Dataset D. This resulted in a significant difference in host overlap of parasites relative to a random expectation. If dropping the minimum records of parasites (thereby introducing more parasites to the analysis) removed the significant effect, we would be worried that this was due to error in host distributions associated with lower sample sizes.

Similarly, if increasing the minimum records removed the significant effect we would be worried that we removed important data that were driving the initial relationship (the same logic can be applied to the comparisons of our data with the MalAvi dataset). Nevertheless, we report that restricting the minimum number of parasite records to 5 (from 10; Dataset D) gives similar results to the original analysis (*Plasmodium* observed mean Bray–Curtis dissimilarty [obs. MBC] = 0.73, p = 0.09, *Haemoproteus* obs. MBC = 0.93, p < 0.01, *Leucocytozoon* obs. MBC = 0.90, p < 0.01); likewise for increasing the minimum number of parasite records to 15 (*Plasmodium* obs. MBC = 0.73, p = 0.10, *Haemoproteus* obs. MBC = 0.94, p < 0.01, *Leucocytozoon* obs. MBC = 0.85, p < 0.01).

In general, we are more worried about sample size affecting results when we fail to reject a null hypothesis. The tests of hypotheses P3–P5 are such cases where we failed to reject null hypotheses and so we re-ran those tests and we report the results here. This should give the reader confidence that our results were not driven by our sample size (number of records) choice. For hypothesis P3, we failed to reject the null hypothesis that mean host specificity of parasites is unrelated to host abundance. If we reduce the parasite minimum number of records to 5 (from 10; Dataset D), the relationships remain non-significant and the estimates of the coefficients are almost identical to the original analysis (*Haemoproteus* Gini–Simpson index of host specificity, β = -0.03, p = 0.38, Rao's QE of host specificity, β= -0.97, p = 0.25; *Leucocytozoon*, [GS] β = -0.04, p = 0.19, [RQE] β = -1.54, p = 0.22). Increasing the parasite minimum number of records to 15 still results in no relationships and similar coefficients (*Haemoproteus* [GS], β = -0.02, p = 0.55, [RQE], β= -0.74, p = 0.38; *Leucocytozoon*, [GS] β = -0.03, p = 0.38, [RQE] β = -0.91, p = 0.42). Similarly for hypothesis P4 (lineage prevalence as a function of host specificity), reducing the parasite minimum number of records to 5 (*Haemoproteus* [GS], β = 0.31, p =0.62, [RQE], β=0.01, p =0.73; Leucocytozoon, [GS] β =-1.04, p =0.28, [RQE] β =-0.03, p =0.24) or increasing it to 15 (*Haemoproteus* [GS], β =-0.16, p =0.84, [RQE], β=-0.01, p =0.60; *Leucocytozoon*, [GS] β =-1.92, p =0.17, [RQE] β =-0.07, p =0.08) does not change the results.

In the manuscript we repeated the test of hypothesis P4 with maximum prevalence and re-ran the models. We examined what happens when we change the parasite minimum number of records to 5 (*Haemoproteus* [GS], β = 0.34, p =0.75, [RQE], β=0.00, p =0.91; *Leucocytozoon*, [GS] β =-0.76, p =0.51, [RQE] β =-0.01, p =0.72) and then to 15 (*Haemoproteus* [GS], β = 0.30, p =0.76, [RQE], β=0.01, p =0.67; *Leucocytozoon*, [GS] β =-1.05, p =0.39, [RQE] β =-0.04, p =0.33). In both cases the results were consistent with the results we obtained with the minimum number of records of 10 from Dataset D.

Hypothesis P5 tested lineage abundance as a function of host specificity. Here we found a positive relationship for *Haemoproteus* and no relationship for *Leucocytozoon*. However, the

positive relationship for *Haemoproteus* was driven by three lineages of the morphospecies *Haemoproteus majoris*. Dropping the minimum number of parasite records to 5 (*Haemoproteus* [GS], β = 0.89, p =0.01, [RQE], β=0.03, p =0.01; *Leucocytozoon*, [GS] β =0.51, p =0.21, [RQE] β =0.02, p =0.26; removing *H. majoris*, [GS], β = -0.04, p =0.91, [RQE], β=0.00, p =0.92) and increasing it to 15 (*Haemoproteus* [GS], β = 0.70, p =0.03, [RQE], β=0.02, p =0.03; *Leucocytozoon*, [GS] β =0.60, p =0.15, [RQE] β =0.02, p =0.28; removing *H. majoris*, [GS], β = -0.20, p =0.39, [RQE], β=-0.01, p =0.25) gives us results that are consistent with the original analysis.

Thus, by our estimation, our results are relatively robust to variation in the sample size (or number of records) that one uses in the analyses.

References

Jovani, R. and Tella, J. L. 2006. Parasite prevalence and sample size: misconceptions and solutions. – Trends Parasitol. 22: 214–218.

Table A1. A list of the host species sampled at Krankesjön from 2013 to 2017 which could be classified into $\log_2$ abundance categories and their sample sizes broken down by age categories (AHY = adult, HY = juvenile, Unknown = could not be aged; data from recaptured individuals were restricted to the first time the individuals were captured because some changed their age category between captures).

Species	Sample size by age			Log_2 Abundance
	AHY	HY	Unknown	
Acrocephalus palustris	14	42	0	5
Acrocephalus schoenobaenus	33	69	0	6
Acrocephalus scirpaceus	36	74	0	7
Aegithalos caudatus	17	10	6	4
Alcedo atthis	5	17	0	1
Anthus trivialis	25	10	1	6
Carduelis cannabina	0	5	0	3
Carduelis carduelis	3	16	0	4
Carduelis chloris	26	13	0	3
Carduelis flammea	19	2	0	4
Carduelis spinus	5	4	0	2
Certhia familiaris	11	12	2	5
Columba palumbus	1	0	0	5
Corvus monedula	45	36	0	5

Dendrocopos minor	1	2	0	1
Dendrocopus major	6	5	0	3
Drycopus martius	2	0	0	2
Emberiza citrinella	32	16	0	6
Emberiza schoeniclus	41	54	0	6
Erithacus rubecula	48	61	0	7
Ficedula hypoleuca	1	8	0	1
Fringilla coelebs	77	17	0	8
Garrulus glandarius	1	0	0	3
Hippolais icterina	38	13	0	5
Hirundu rustica	1	0	0	3
Lanius collurio	15	3	0	2
Locustella fluviatilis	1	1	0	0
Locustella naevia	9	3	0	3
Luscinia luscinia	25	8	0	5
Motacilla alba	1	6	0	3
Muscicapa striata	4	0	0	4
Panurus biarmicus	0	11	11	5
Parus ater	5	2	0	4
Cyanistes caeruleus	59	115	0	8
Parus major	59	122	0	7
Poecile palustris	47	36	0	6
Passer montanus	4	2	10	3
Phasianus colchicus	0	1	0	4
Phoenicurus phoenicurus	8	17	0	3
Phylloscopus collybita	68	76	0	7
Phylloscopus sibilatrix	18	0	0	3
Phylloscopus trochilus	105	104	0	9
Pica pica	0	2	0	3
Picus viridis	0	1	0	2
Prunella modularis	33	59	0	6
Pyrrhula pyrrhula	2	0	0	1
Regulus regulus	15	15	0	4
Remiz pendulinus	0	1	0	1
Saxicola rubetra	15	25	0	5
Sitta europaea	11	13	6	5
Sturnus vulgaris	13	28	0	5
Sylvia atricapilla	117	53	0	8

Sylvia borin	62	36	0	7
Sylvia communis	60	71	0	6
Sylvia curruca	11	21	0	3
Troglodytes troglodytes	22	24	1	5
Turdus merula	25	11	0	6
Turdus philomelus	22	5	0	5
Turdus pilaris	1	0	0	4
Turdus viscivorus	1	0	0	1

Table A2. Spearman rank correlations between the prevalence and diversity (Gini–Simpson index of parasite lineages) of each of the three parasite genera among host species sampled at Krankesjön.

	ρ	p
<u>Prevalence</u>		
Plasmodium versus *Haemoproteus*	0.17	0.298
Plasmodium versus *Leucocytozoon*	0.28	0.084
Haemoproteus versus *Leucocytozoon*	0.29	0.074
<u>Diversity</u>		
Plasmodium versus *Haemoproteus*	-0.12	0.457
Plasmodium versus *Leucocytozoon*	0.16	0.329
Haemoproteus versus *Leucocytozoon*	0.15	0.355

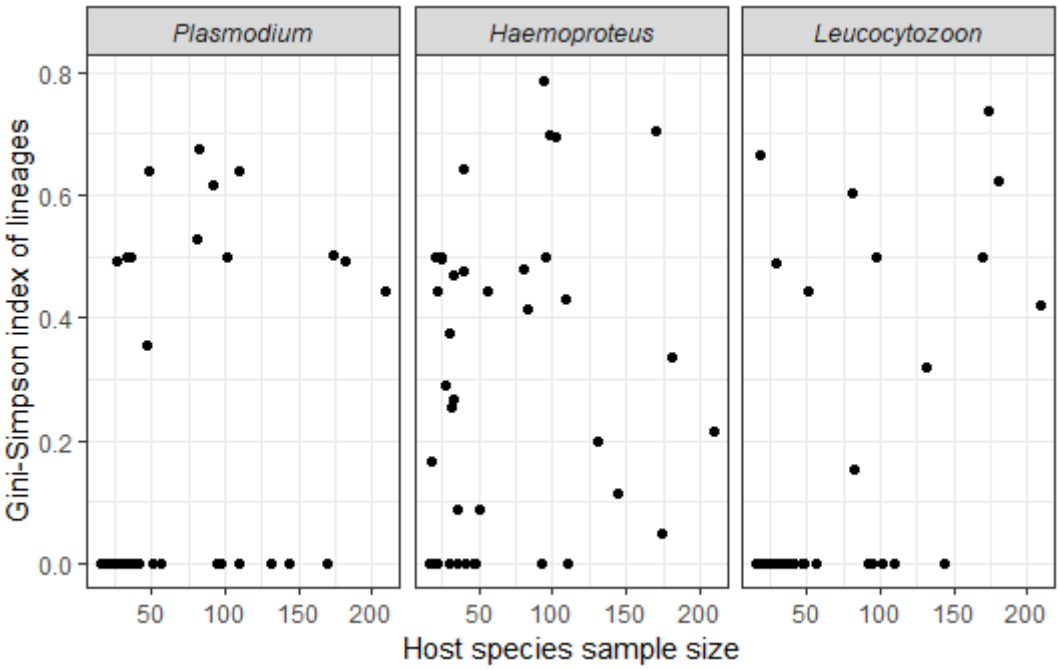

Figure A1. Relationship between the Gini-Simpson diversity of lineages infecting each host species and the sample size of those host species at Krankesjön. Correlations were significant, but weak for *Plasmodium* ($r = 0.34$, $p = 0.036$) and *Leucocytozoon* ($r = 0.51$, $p < 0.001$), and not significant for *Haemoproteus* ($r = 0.16$, $p = 0.323$). Points in the graph are species.

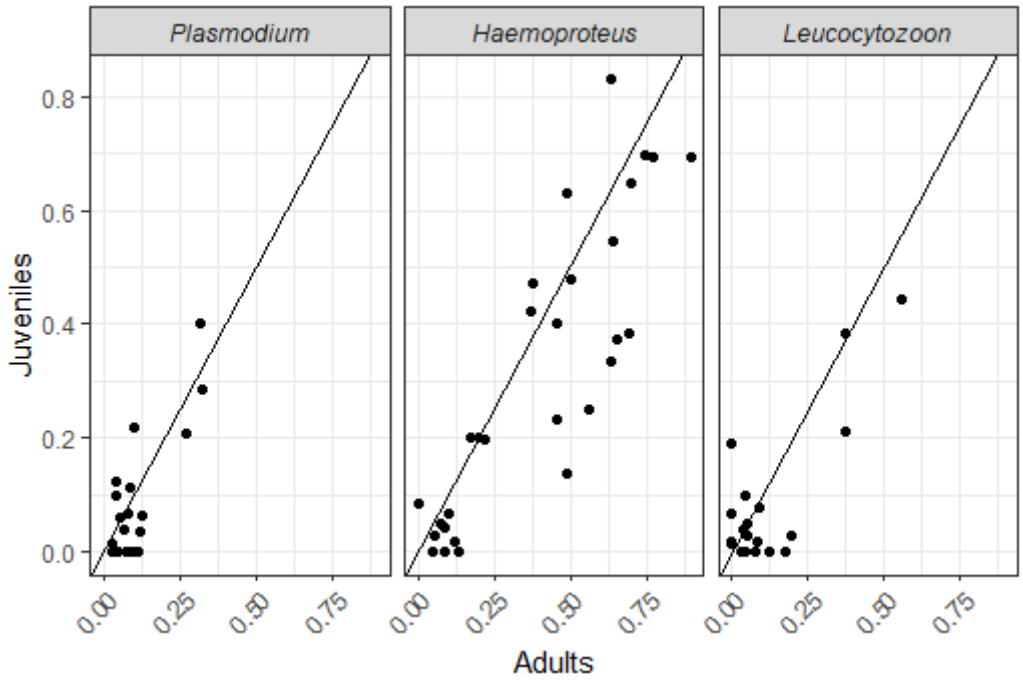

Figure A2. Relationship between adults and juveniles of host species at Krankesjön in terms of parasite prevalence. A line of slope one intersecting the origin of the graph is shown.

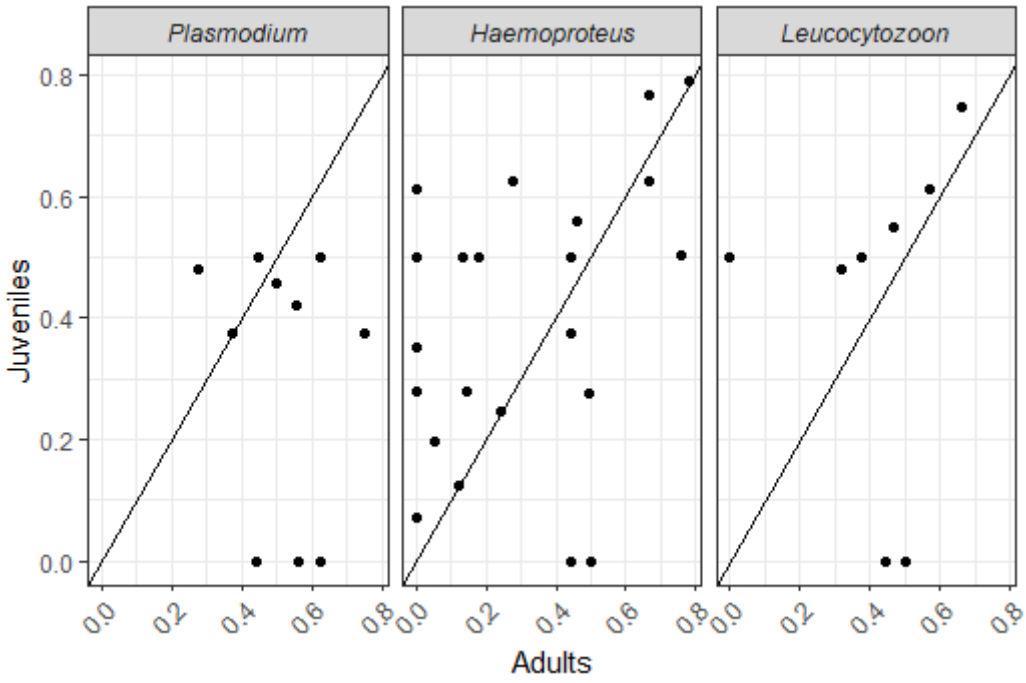

Figure A3. Relationship between adults and juveniles of host species at Krankesjön in terms of the diversity of parasite lineages they were infected by. A line of slope one intersecting the origin of the graph is shown.

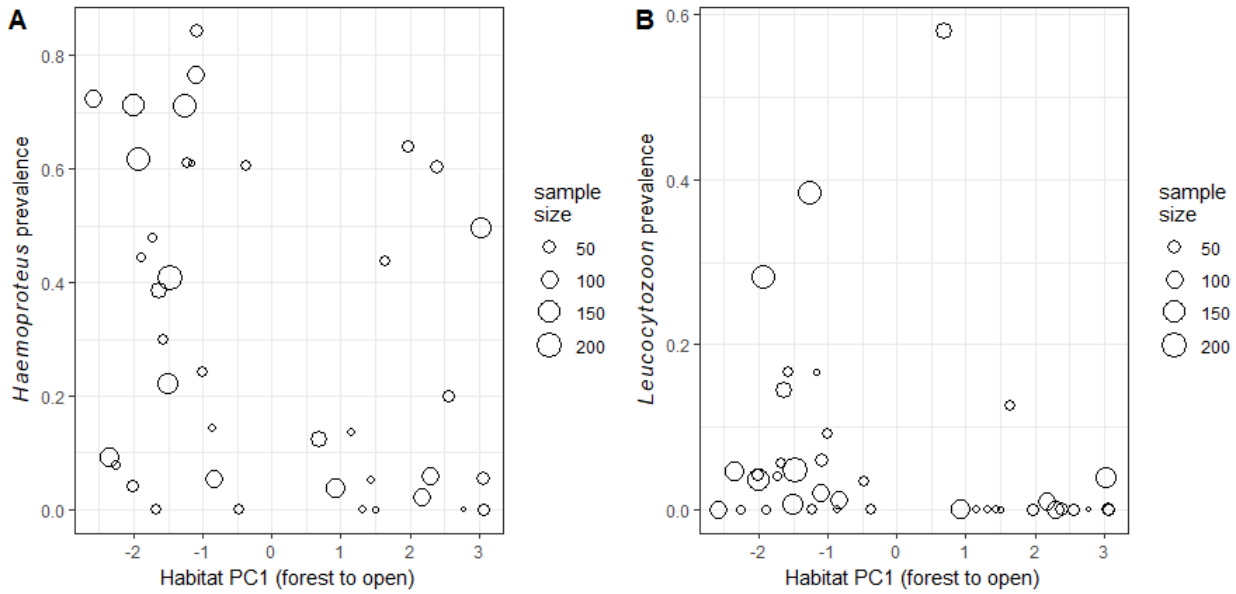

Figure A4. The relationship between the prevalence of *Haemoproteus* (A) and *Leucocytozoon* (B) parasites of host species at Krankesjön and habitat PC1 (the habitat preferences of those host species corresponding to habitat openness).

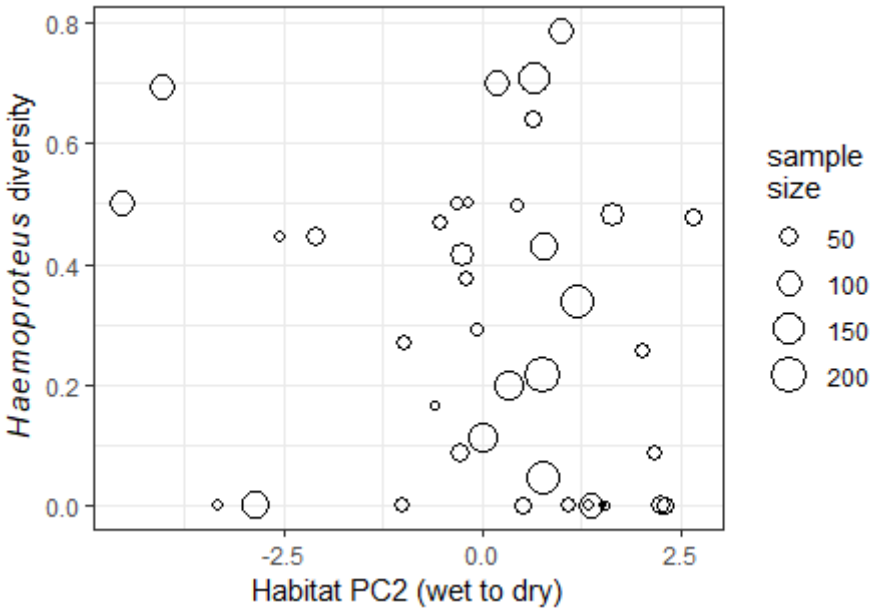

Figure A5. Relationship between the diversity of *Haemoproteus* parasites (Gini–Simpson index) in host species of Krankesjön and habitat PC2 corresponding to the habitat preferences of those host species (ranging from wet to dry habitat).

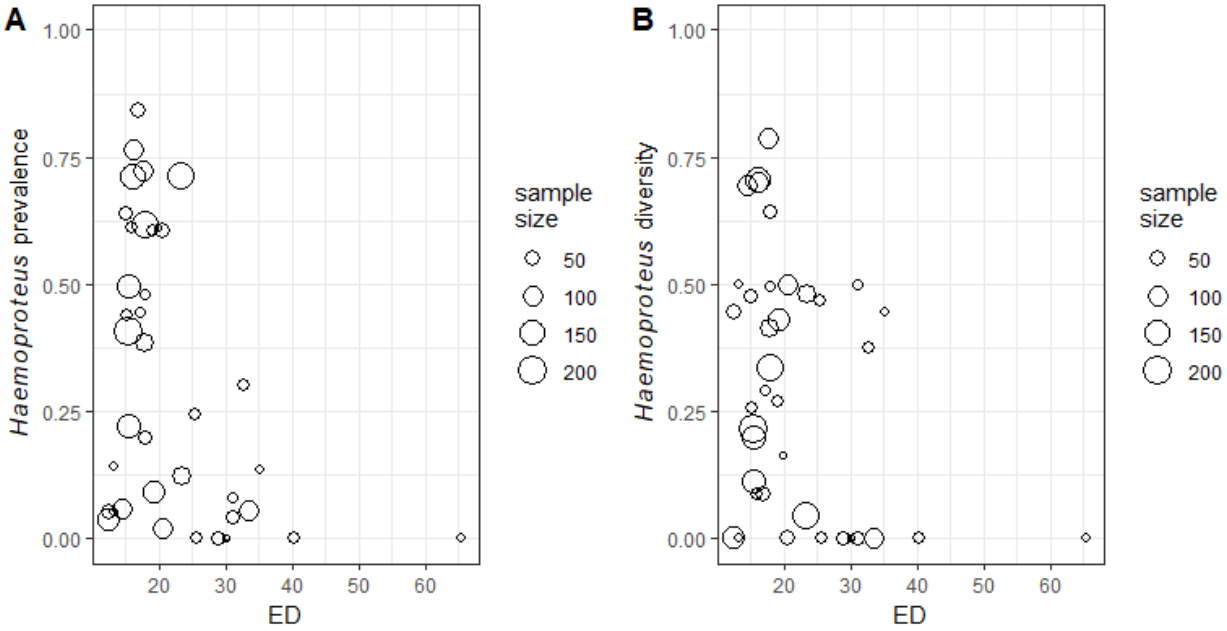

Figure A6. Relationship between the prevalence (A) and diversity (B) of *Haemoproteus* parasite lineages (the latter calculated as a Gini–Simpson index) of host species at Krankesjön and the evolutionary distinctiveness ('ED', low ED corresponds to species with many close relatives and high ED corresponds to species with few close relatives) of those host species.

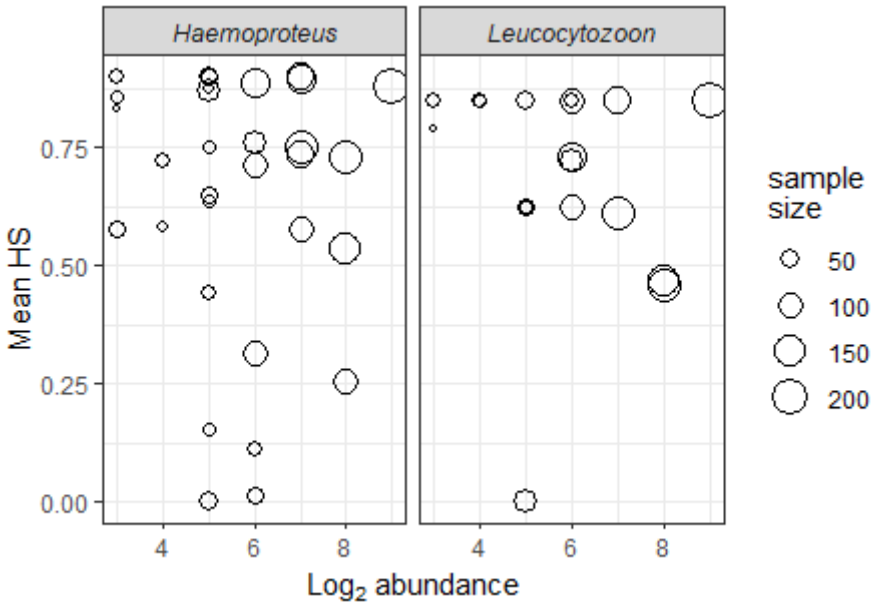

Figure A7. Relationship between the mean host specificity (Gini–Simpson index) of each host species' parasite lineages at Krankesjön and the abundance of those host species. Points are host species.

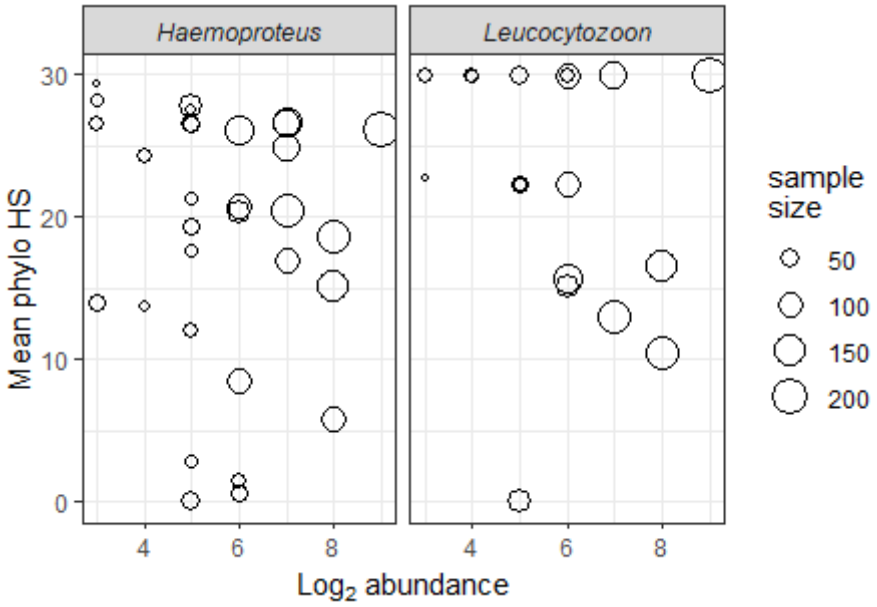

Figure A8. Relationship between the mean host specificity (Rao's quadratic entropy) of each host species' parasite lineages at Krankesjön and the abundance of those host species. Points are host species.

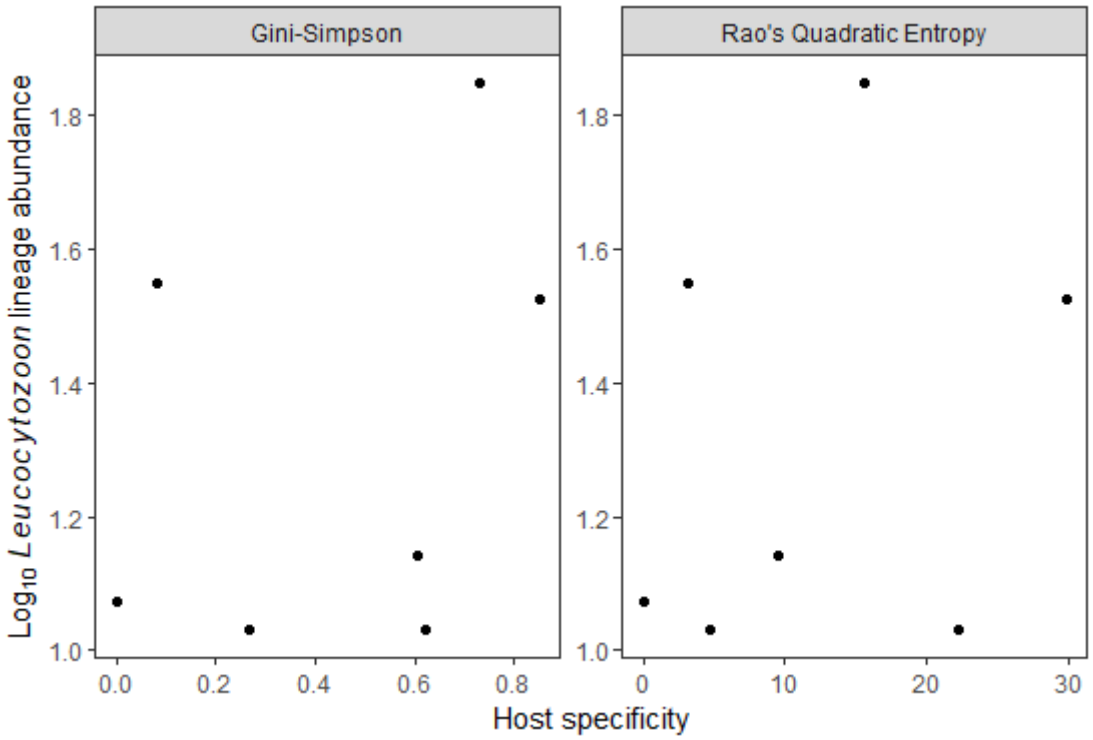

Figure A9. Relationship between the abundance (log_{10} transformed) of *Leucocytozoon* lineages at Krankesjön and host specificity represented as both a Gini–Simpson index and Rao's quadratic entropy.