Supplemental Material

Predicting parasite community composition of individual hosts using joint species distribution models

Tad Dallas[a,b,*], Anna-Liisa Laine[a,c], and Otso Ovaskainen[a,d]

[a] Organismal and Evolutionary Biology Research Programme, P.O. Box 65, 00014 University of Helsinki, Finland

[b] Department of Biology, Louisiana State University, Baton Rouge, LA, USA

[c] Department of Evolutionary Biology and Environmental Studies, University of Zürich, CH-8057 Zürich, Switzerland

[d] Centre for Biodiversity Dynamics, Department of Biology, Norwegian University of Science and Technology. N-7491 Trondheim, Norway

[*] tad.a.dallas@gmail.com

Host tissue specificity

Ectoparasitic species were only found infecting the host coat, while parasite species infecting the host gastrointestinal tract exhibited some variation in their dominant host tissue (Figure S1).

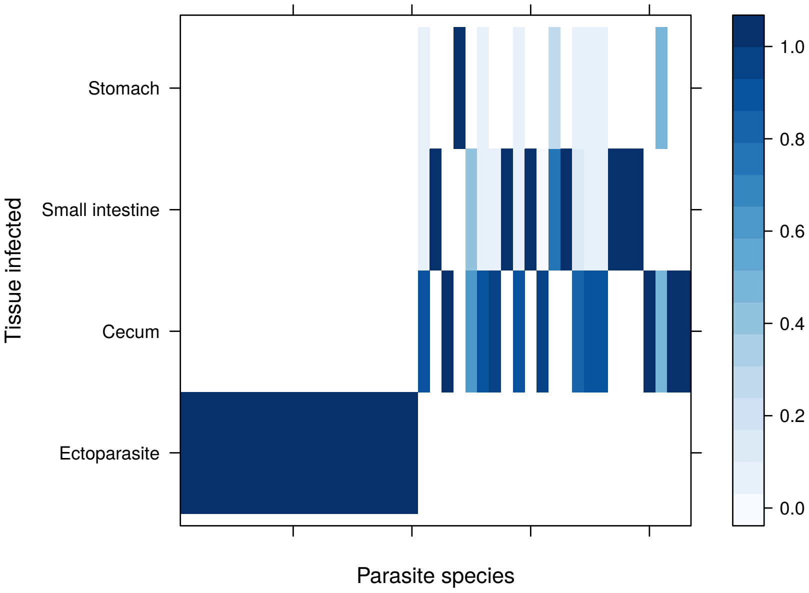

Figure S1: Parasite host tissue utilization, standardized as the fraction of times the parasite was found in each host tissue (indicated by color).

The relationship between measures of model performance

Tjur's R^2 and AUC were used to quantify model performance in the main text. To validate that these measures were at least measuring model performance in a similar manner, we examine the relationship between them here. We find a strong positive relationship between the two measures of model performance, despite the differences in quantification. That is, AUC is based solely on ranks, and is a measure of model discrimination (i.e., the ability of the model to rank positive cases from negative cases in the test set). Meanwhile, Tjur's R^2 is a pseudo-R^2 measure which uses the model-predicted probabilities of occurrence in the host-parasite matrix, quantifying the difference between the probability of occurrence and the probability of absence for each potential occurrence (i.e., a parasite species on a host individual).

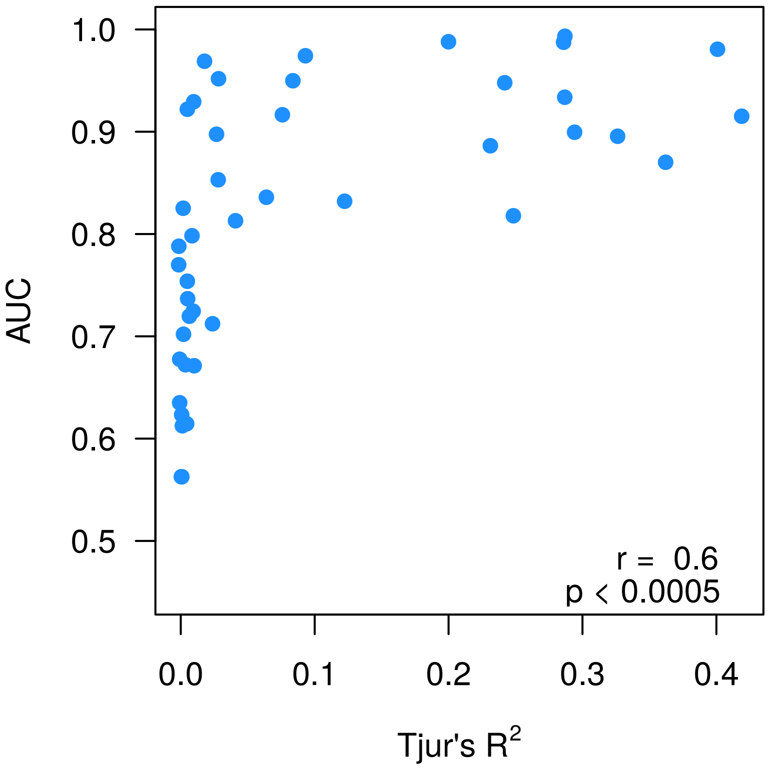

Figure S2: The relationship between Tjur's R^2 and AUC, where each point corresponds to a host species. Reported statistics are based on a Spearman's rank correlation.

Model structure and random effects

In the main text, we examine the performance of models which tacitly assume that random effects corresponding to geographic, temporal, and host variation influences the parasite community of a given host individual. This is supported by the variance partitioning analysis. Here, we further support the importance of the inclusion of these random effects by comparing a set of five models.

The models are as follows:

- **Model 1**: No random effects.
- **Model 2**: Excluding the host individual random effect, but including other random effects.
- **Model 3**: Including host individual random effect, but exclude other random effects.
- **Model 4**: Including all random effects.
- **Model 5**: Including species associations in model predictions (i.e., conditional predictions made for subset of parasite species conditional on known occurrences of the non-focal species)

All models described above contained host sex as a fixed effect, and were fit using the procedure described in the main text; 4 MCMC chains with $100 \cdot$ thin iterations used for burn-in and $200 \cdot$ thin iterations for the actual sampling, thinning value of 100. Models were 5-fold cross validated, as in the main text. We find that the inclusion of the host individual random effect is not nearly as important to model performance as the other random effects, but that the full model performed best, suggestive of a clear benefit of the incorporation of random effects in the HMSC framework for the prediction of ecological communities (Table S1). The lack of importance of the individual level random effect is simply due to the way in which we estimate model performance, as 5-fold cross validation ablates the influence of the individual level random effect on model performance.

Table S1: Model performance examining the set of five models described above, which varied in their inclusion of random effects (RE) and effect of the individual (Ind). Model 5 is identical to model 4 in structure, but uses the species associations to conditionally predict community composition.

Model	Ind	RE	$A\bar{U}C$	SD_{AUC}	$Tj\bar{u}rR^2$	$\mathrm{SD}_{T}jurR^2$	$R\bar{M}SE$	SD_{RMSE}
1			0.58	0.11	0.00	0.00	0.12	0.10
2		✓	0.82	0.14	0.10	0.14	0.11	0.08
3	✓		0.60	0.10	0.00	0.00	0.12	0.10
4	✓	✓	0.81	0.15	0.10	0.14	0.11	0.08
5	✓	✓	0.82	0.09	0.09	0.08	0.11	0.09

We further explore the influence of each random effect individually, by training models (as described above) considering models including host sex as a fixed effect, but including each random effect independently. Residual covariance matrices (Ω_i) as a function of each random effect as provided in Figure S3. Variation existed as a function of sampling year, site, and species, suggesting that parasite distributions among host individuals varied geographically and temporally. Residual covariance matrix values (Ω_ij) falling below the threshold of 90% posterior probability were removed.

Further, we explored the influence of each random effect on the subsequent parasite association (ω) matrix. We trained models containing each random effect in isolation, finding that incorporating host species as a random effect resulted in the greatest increase in model performance out of any single random effect (Table S2).

Table S2: Model performance for a set of models incorporating a single random effect, demonstrating the clear importance of including random effects in terms of model performance, as well as the large effect of incorporating information on host species. Each model was 5-fold cross-validated and trained as in the main text, where we used 100 · thin iterations used for burn-in and 200 · thin iterations for the actual sampling.

Random effect	$\bar{AUC}$	SD$_{A}UC$	$TjurR^2$	SD$_{T}jurR^2$	$\bar{RMSE}$	SD$_{RMSE}$
Individual	0.57	0.11	-0.00	0.00	0.12	0.10
Year	0.60	0.13	0.01	0.02	0.12	0.10
Site	0.67	0.15	0.02	0.04	0.12	0.10
Season	0.60	0.10	0.00	0.01	0.12	0.10
Species	0.76	0.15	0.06	0.10	0.11	0.09

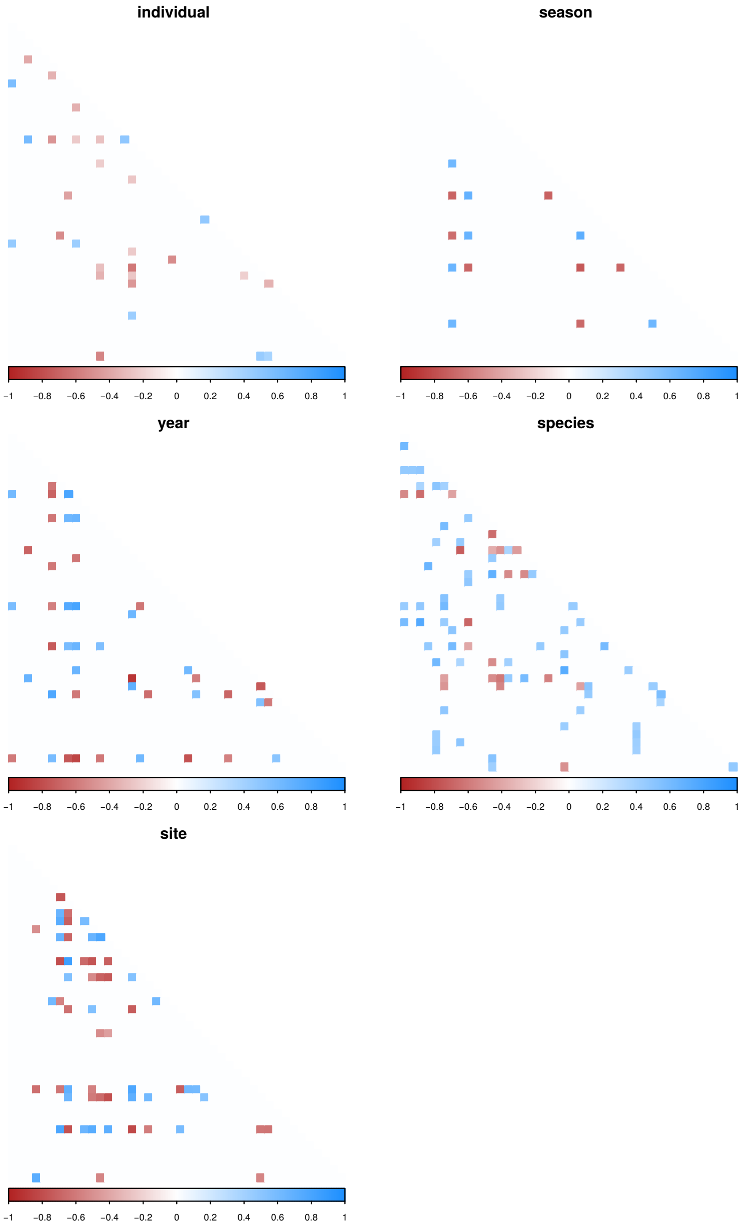

Figure S3: Residual covariance (Ω) matrices for each random effect. *Individual* is equivalent to main text Figure 4, while the other matrices are presented here for completeness.

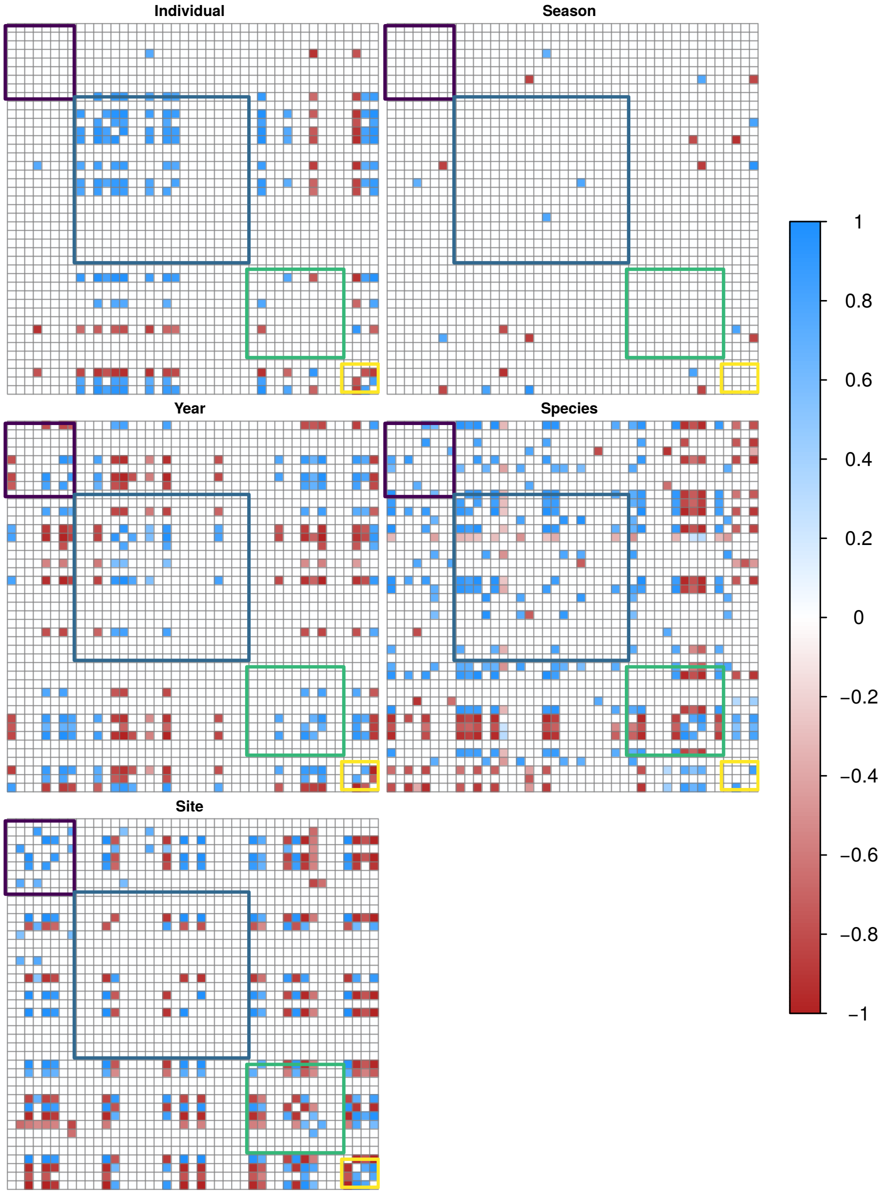

Figure S4: Residual covariance (Ω) matrices for each random effect when only that random effect was included in the model.

Season as a fixed effect

In the main text, we treated season as a random effect, though random effects tend to have fewer random levels. Here, for thoroughness, we examined the relative effect on model performance of including season as a fixed effect instead of a random effect. We find little evidence of a difference in model performance when estimating species richness between models incorporating season as a random effect (Table S3).

Table S3: Model performance considering season as a random effect (as in the main text) or a fixed effect. Each model was 5-fold cross-validated and trained as in the main text, where we used 100 · thin iterations used for burn-in and 200 · thin iterations for the actual sampling.

Season	$A\bar{U}C$	SD$_AUC$	$Tj\bar{u}rR^2$	SD$_TjurR^2$	$R\bar{M}SE$	SD$_{RMSE}$
Random	0.812	0.15	0.099	0.14	0.107	0.08
Fixed	0.801	0.17	0.099	0.13	0.107	0.08

Including parasite species only found in host feces

In the main text, we excluded parasite species that were only found in host feces. Here, we include them to examine how model accuracy and parasite association estimation changes as a function of their inclusion. Including these host individuals infected by parasite species only found in host feces increases the number of host individuals from 1347 to 2558, and increases the number of parasite species considered from 43 to 65 (see Table S4 for host individual distributions among the study sites).

We find numerous similarities between analyses in terms of variable importance (Figure S5), model performance (Figure S6), and parasite associations (Figure S7).

Table S4: Number of host individuals sampled for parasite species at each of the six habitats and three habitat types (grassland, larrea, and woodland). Two of the sites (Five points and Rio Salado) contain both grassland and larrea habitat types.

	Grassland		Larrea		Woodland		
Host species	Five points	Rio Salado	Five points	Rio Salado	Sepultrua	Two-twenty-two	Total
Ammospermophilus interpres	0	0	0	0	2	0	2
Chaetodipus intermedius	0	0	0	0	7	53	60
Dipodomys merriami	9	106	316	236	0	104	771
Dipodomys ordi	102	201	1	25	0	0	329
Dipodomys spectabilis	75	0	68	5	0	1	149
Eutamias dorsalis	0	0	0	0	0	7	7
Eutamias quadrivittatus	0	0	0	0	20	0	20
Neotoma albigula	1	101	8	31	29	35	205
Neotoma micropus	2	0	5	1	0	0	8
Onychomys arenicola	34	6	34	6	0	0	80
Onychomys leucogaster	0	64	0	32	0	0	96
Perognathus flavus	1	94	2	21	0	8	126
Perognathus flavescens	125	1	41	36	16	38	257
Peromyscus boylii	0	0	0	0	6	7	13
Peromyscus difficilis	0	0	0	0	1	1	2
Peromyscus eremicus	0	1	7	2	0	0	10
Peromyscus leucopus	2	35	0	40	0	4	81
Peromyscus maniculatus	0	3	0	0	0	0	3
Peromyscus truei	7	22	1	34	163	13	240
Reithrodontomys megalotis	3	18	5	28	1	0	55
Reithrodontomys montanus	3	2	1	10	0	0	16
Sigmodon hispidus	0	1	0	2	0	0	3
Spermophilus spilosoma	15	7	1	2	0	0	25
Total	379	662	490	511	245	271	**2558**

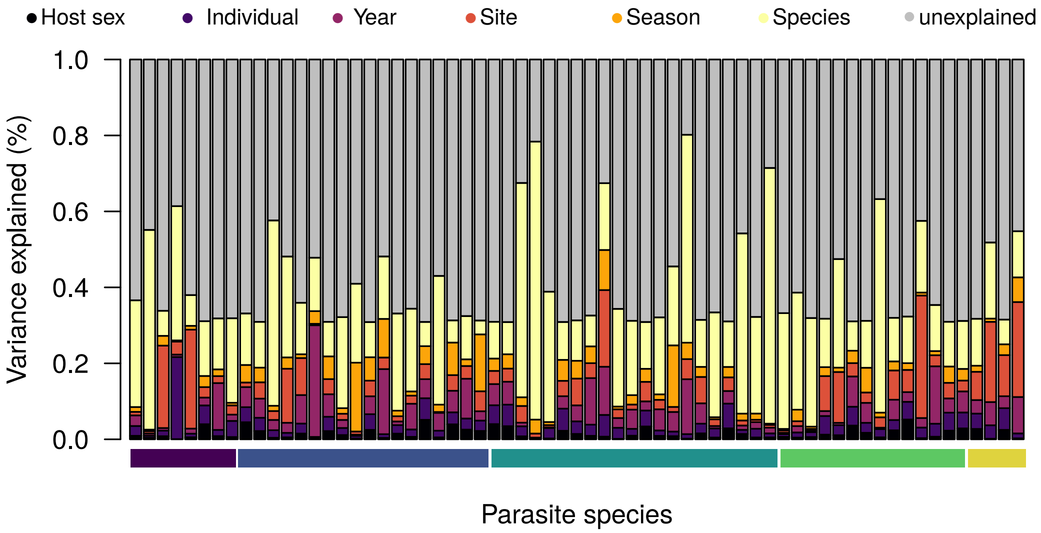

Figure S5: Variance partitioning plot showing differential contributions of geography and host traits to parasite species distributions among host individuals.The colored bar at the bottom identifies parasite species by the dominant host tissue they are found to infect (from left to right; cecum, host pelage (ectoparasite), feces, small intestine, and stomach). Unexplained variance is proportional to Tjur's R^2 of the trained model.

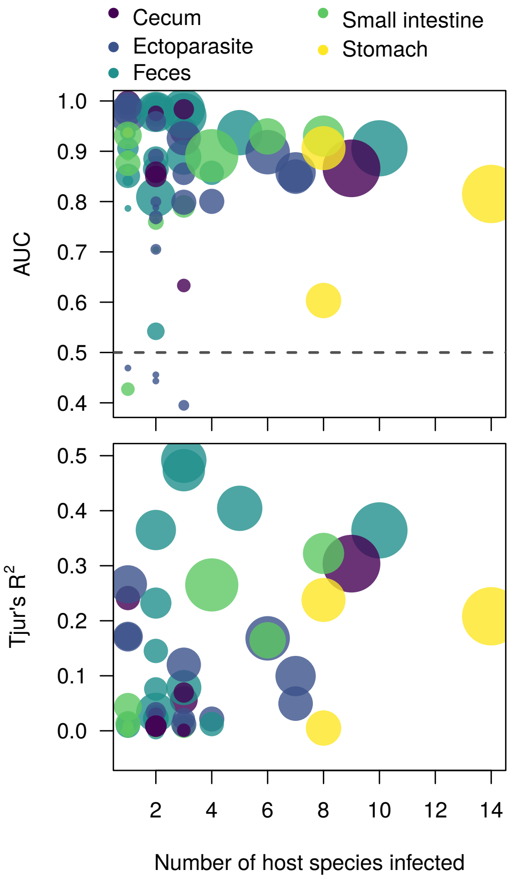

Figure S6: AUC and Tjur's R^2 as a function of the number of host species infected. Each point represents a parasite species, and point size is proportional to the log number of individuals that each parasite species infects.

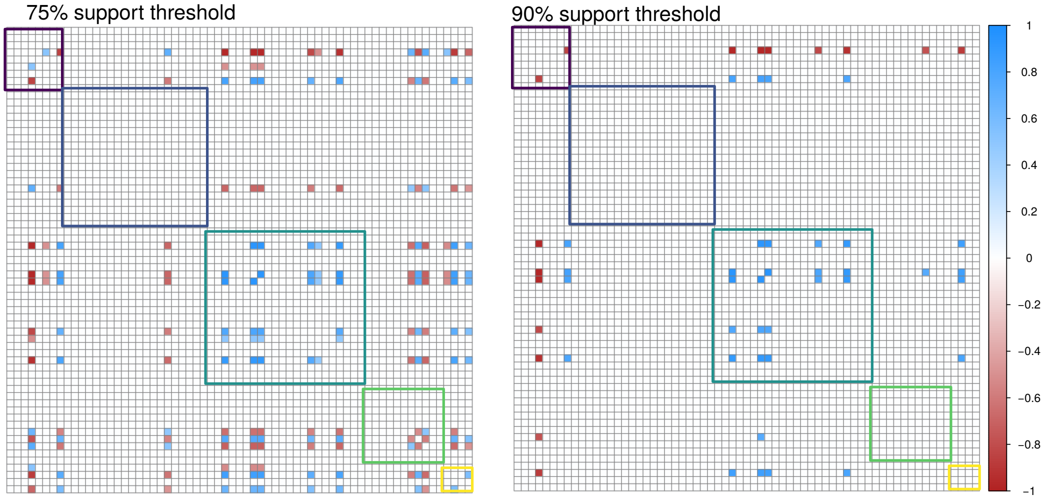

Figure S7: Parasite-parasite associations after controlling for the effects of geographic and host trait covariates for two different support thresholds; 75% (left panel) and 90% (right panel) posterior probability support. Parasite species are ordered based on the host tissue they predominantly infect, with colored boxes indicating parasite-parasite associations when the same host tissue is infected (cecum in purple; ectoparasites in blue; feces in green; small intestine in light green; stomach in yellow).

Examining model convergence

We examined model convergence by increasing the length of each MCMC chain and exploring differences in estimated association (Ω) matrices. There was no large differences in estimated association matrices as a function of chain length (Figures S8 - S12), especially after the chain length exceeded 10,000.

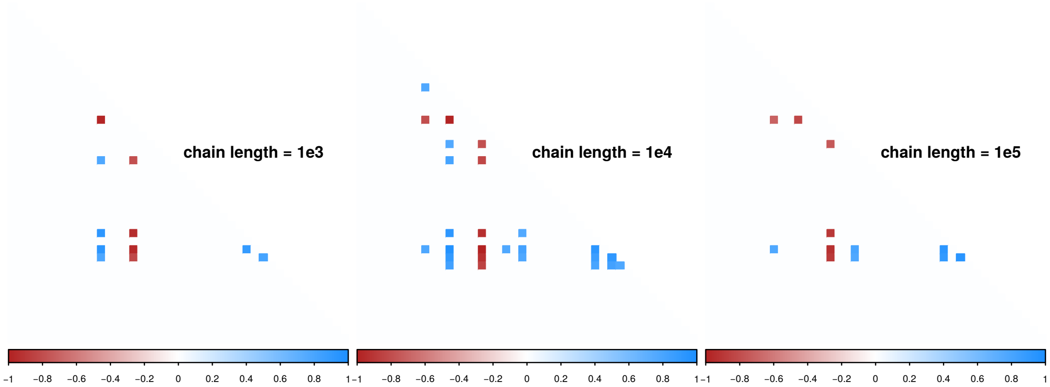

Figure S8: Association (Ω) matrices examining the influence of MCMC chain length ($1{\times}10^3$ - $1{\times}10^5$) on estimation of residual covariance among parasite species as a function of host individual after accounting for the other random effects.

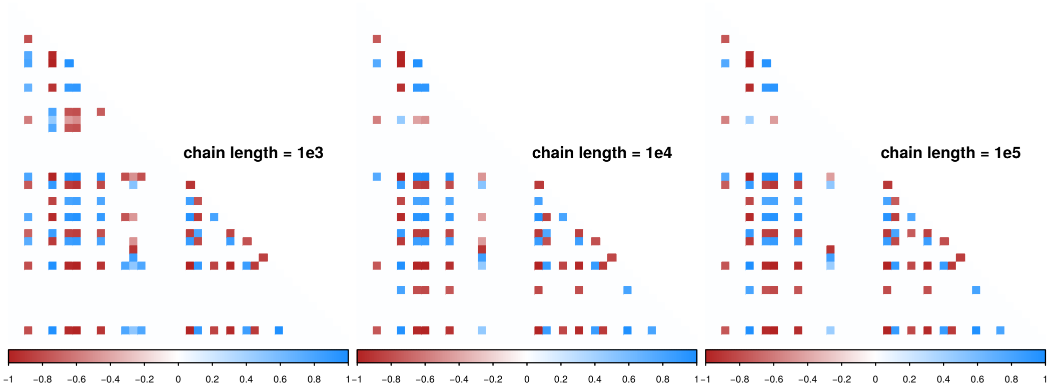

Figure S9: Association (Ω) matrices examining the influence of MCMC chain length (1×10^3 - 1×10^5) on estimation of residual covariance among parasite species as a function of sampling year after accounting for the other random effects.

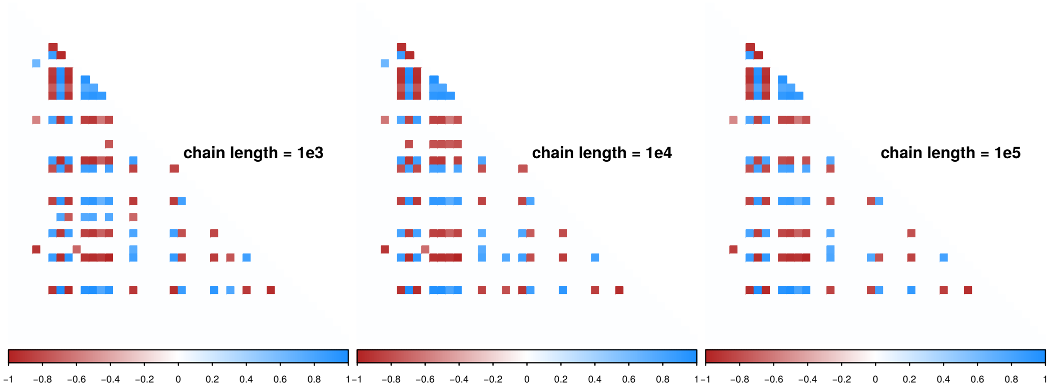

Figure S10: Association (Ω) matrices examining the influence of MCMC chain length (1×10^3 - 1×10^5) on estimation of residual covariance among parasite species as a function of sampling site after accounting for the other random effects.

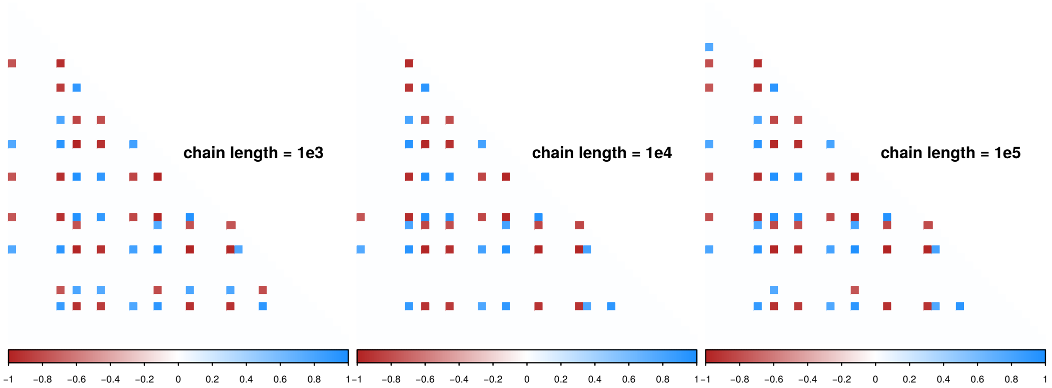

Figure S11: Association (Ω) matrices examining the influence of MCMC chain length (1×10^3 - 1×10^5) on estimation of residual covariance among parasite species as a function of sampling season after accounting for the other random effects.

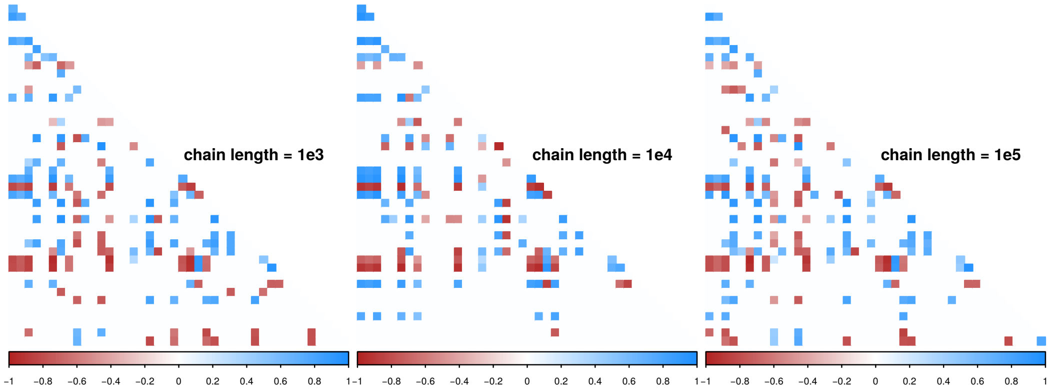

Figure S12: Association (Ω) matrices examining the influence of MCMC chain length (1×10^3 - 1×10^5) on estimation of residual covariance among parasite species as a function of host species after accounting for the other random effects.