Submission type: Article - Discoveries

Title: Uncovering the genomic basis of infection through co-genomic sequencing of hosts and parasites

Authors: Eric Dexter[1]*, Peter D. Fields[1], and Dieter Ebert[1]

Affiliations:

[1]Department of Environmental Sciences, The University of Basel; Basel, 4051, Switzerland

*Corresponding author email: eric.dexter@unibas.ch

Short title: Co-genomic sequencing of hosts and parasites

Abstract:

Understanding the genomic basis of infectious disease is fundamental objective in coevolutionary theory with relevance to healthcare, agriculture, and epidemiology. Models of host-parasite coevolution often assume that infection requires specific combinations of host and parasite genotypes. Coevolving host and parasite loci are therefor expected to show associations that reflects an underlying infection/resistance allele matrix, yet little evidence for such genome-to-genome interactions has been observed among natural populations. We conducted a study to search for this genomic signature across 258 linked host (*Daphnia magna)* and parasite (*Pasteuria ramosa)* genomes. Our results show a clear signal of genomic association between multiple epistatically-interacting loci in the host genome, and a family of genes encoding for collagen-like protein in the parasite genome. These findings are supported by laboratory-based infection trials, which show strong correspondence between phenotype and genotype at the identified loci. Our study provides clear genomic evidence of antagonistic coevolution among wild populations.

Keywords: Coevolution, parasite, GWAS, Daphnia, Pasteuria, genome-to-genome, co-genomics

INTRODUCTION

Parasites (including pathogens) infect nearly all forms of life, acting as a fundamental biological force that shapes the evolution of their hosts and drives the dynamics of entire communities (Poulin 2011; Schmid-Hempel 2021). The evolutionary processes among host and parasite are tightly interconnected: As a host population evolves in response to the threat caused by its parasites, so too does the parasite evolve in response to its host. This antagonistic coevolution may be responsible for much of the genetic diversity found in natural populations, promoting local adaptation and population differentiation (Morand 2015; Ebert and Fields 2020). Various models of host-parasite coevolution have been developed to accommodate the fact that phenotypic traits under selection may have different genomic architectures. Under highly specific models of host-parasite interactions, such as the matching-allele and the gene-for-gene models, infections are only possible between specific combinations of host and parasite genotypes (Agrawal and Lively 2002). Parasites will therefore experience high fitness when compatible alleles are common in the host population, and low fitness when such alleles are rare, thus driving a time-lagged cycle of negative-frequency dependent selection which is sometimes known as the Red Queen model of coevolution (Ebert and Fields 2020).

Red Queen evolutionary dynamics are predicted to yield several diagnostic characteristics in both host and parasite genomes. As negative-frequency dependent selection promotes the maintenance of genetic diversity, affected loci are expected to exhibit locally elevated nucleotide diversity relative to other genomic regions (Ebert and Fields 2020). This is because selection is tied to allele frequencies, with common variants being selected against. Highly specific models of host-parasite interactions also predict that certain combinations of host and parasite alleles are unlikely to be found in infected hosts, even if both alleles occur in the population at large. Coevolving host and parasite loci are therefore expected to show a statistical pattern of association that reflects the underlying infection/resistance allele matrix (Ebert and Fields 2020). The non-random association of alleles across the genomes of two interacting species can be considered as a form of interspecies linkage disequilibrium, more succinctly "interlinkage." Interlinkage is a predicted consequence of genomic interactions between hosts and parasites and thus direct evidence for ongoing coevolutionary interactions of the antagonists. Despite arising as a fundamental prediction of host-parasite coevolution, there is little empirical evidence of interlinkage among naturally coevolving populations.

Here we present clear evidence of interlinkage among natural populations of the planktonic crustacean *Daphnia magna* and its endoparasitic bacterium *Pasteuria ramosa* (Firmicutes).

Daphnia magna is a common planktonic microcrustacean that inhabits standing fresh and brackish waters across most of the northern hemisphere. Wild populations of *D. magna* are often heavily parasitized by a wide variety of bacteria, virus, and microsporidian species, including the endospore-forming bacteria *Pasteuria ramosa* (Ebert 2005). *Daphnia magna* ingest planktonic spores of *P. ramosa* during filter feeding, which enter the digestive tract along with food and other particulate matter and may attach to the gut lining. After attachment and penetration of the host cuticle, *P. ramosa* rapidly proliferate in the body cavity of the host (Ebert et al. 1996). During *P. ramosa* infections, the normally transparent host body cavity is completely filled with opaque spores which will be released into the environment upon death of the host. *P. ramosa* infection spreads exclusively through horizontal transmission from host to host.

The *D. magna* / *P. ramosa* infection process appears to be governed by highly specific interactions, such that *D. magna* genotypes (aka. clones) are either completely susceptible or completely resistant to particular strains of *P. ramosa* (Bento et al. 2017; Bento et al. 2020). This binary outcome is strongly linked to the attachment of the parasite to the host cuticle. This highly-specific attachment step is hypothesized to be governed by interactions between *P. ramosa* surface proteins and highly-polymorphic extracellular proteins at the target tissue of the host (Andras et al. 2020). Extensive attachment testing with a wide range of host and parasite isolates has shown that susceptibility is almost perfectly correlated with attachment, with little to no variation due to environmental or host condition (Duneau et al. 2011).

Previous studies have shown that *D. magna* resistance to *P. ramosa* attachment is governed by multiple epistatically interacting loci distributed across the *D. magna* genome. Genetic crosses of *D. magna* clones with differing *P. ramosa* resistance phenotypes have thus far identified six resistance loci: the A, B, C, D, E, and F loci. The A, B, and C loci cluster tightly together on chromosome 4 to form the ABC supergene, which shows major structural polymorphism between haplotypes, including long stretches of nonhomologous sequences that appear to be non-recombining (Bento et al. 2017). The F locus is closely located to the ABC cluster, while the D and the E loci are distributed elsewhere across the *D. magna*

genome, but epistatic interactions between combinations of these loci have been described (Bento et al. 2017; Ameline et al. 2021).

The loci which govern infection in the *P. ramosa* genome are less well understood, but there are indications that a highly expanded family of genes encoding collagen-like proteins may be a determinant of infection outcome. Collagen-like proteins are known to be strongly tied to the infection process in a number of bacteria species by mediating attachment to host cells (Qiu et al. 2021). Pasteuria collagen-like (PCL) genes are widely distributed across the *P. ramosa* genome in triplet clusters, potentially the consequence of repeated duplications of a highly mobile transposable element (McElroy et al. 2011). A previous study found that variation at one of these PCLs correlated perfectly with the ability of certain *P. ramosa* isolates to infect a specific *D. magna* genotype (Andras et al. 2020). Given that *P. ramosa* spores are coated with a thick layer of fibers hypothesized by be composed of collagen-like proteins (Andras et al. 2020), and that spore attachment to any of several sites on the host cuticle appears to be the primary determinant of infection outcome(Duneau et al. 2011; Fredericksen et al. 2021), these PCL genes are promising candidates for sites of interlinkage with host resistance genes.

Previous studies of the *D. magna* / *P. ramosa* system have examined loci governing the infection process through experimental contexts which allow for genomic variation only for the host or for the parasite. Although multiple resistance genes in the host and one infectivity gene in the parasite are known, the manner in which these genes interact with each other is not yet understood. Such interactions are at the core of matching-allele and gene-for-gene models, thus forming the most fundamental aspect of coevolutionary dynamics. A number of different statistical approaches have been developed to detect signals of coevolution, both at the genomic level, and the phylogenetic level (Legendre et al. 2002; Carlson et al. 2008; Wang et al. 2018; Märkle et al. 2021). The co-GWAS approach is an extension of single-phenotype GWAS (Genome Wide Association Study), which simultaneously considers genetic information from host and parasite. Through the combination of clinical data and genome-scale SNP arrays, co-GWAS has been used to identify genomic variants of clinical significance for HIV infection (Bartha et al. 2013), Hepatitis C (Ansari et al. 2017), and pneumococcal meningitis (Lees et al. 2019). In order to identify potentially interlinked loci among our system, we conducted a co-genome wide association study (co-GWAS) using wild collected *D. magna* and their naturally associated *P. ramosa* parasites.

RESULTS

Sequencing and mapping

We sequenced 258 *D. magna* naturally infected with *P. ramosa* collected during an epidemic across one season from a single pond, and separated the resulting reads by mapping against reference genomes for the host and parasite. Overall rates of contamination were low for whole-organism sequencing, with 92.8 % (11.0 SD) of reads mapping to either the host or parasite genome (S1 Fig). Mean depth of coverage was 31.0 (SD = 13.0) for *D. magna* and 318 (SD = 217) for *P. ramosa*. This difference in mean coverage reflects the smaller genome size of the parasite, which is approximately 2 Mb versus the approximately 230 Mb host genome. Ten of the samples showed low coverage for either the host or parasite genome and were excluded from further analysis (S1 Fig).

Genomic variation

We identified 21,933 variant positions (SNPs, indels, and SNP/indel combinations) in the *P. ramosa* sequence data. These variants were filtered for base and mapping quality, excessive read depth, and bi-allelic status, resulting in 10,310 variant positions retained for downstream analysis. For the much larger *D. magna* genome, an initial call set of 4,142,757 variants was reduced to 2,211,993 after identical filtering parameters were applied. Population-wide nucleotide diversity (π) was 0.26 for *D. magna* with per sample missingness rates < 1 %. Estimates of genetic diversity within *P. ramosa* samples varied dramatically, likely reflecting that some hosts had been infected by multiple *P. ramosa* genotypes, while others were subject to single infections (Fig 1A). Eight samples were monomorphic at nearly all sites, thus indicating the presence of only a single haploid *P. ramosa* genotype. The remaining samples showed variable rates of *P. ramosa* polymorphism with extreme cases showing variation at up to 50 % of segregating sites, suggesting multiple infections by dissimilar genotypes (Fig 1A).

Population structure

We examined host and parasite samples for hidden population structure using principal component analysis (PCA) and Discriminant Analysis of Principle Components (DAPC). The *D. magna* samples showed no discernable population structure, matching our expectations for a sexually reproducing population contained within a small pond. The strongest principal component for *D. magna* explained 3.5 % of total genomic variation, with subsequent PCs

136 explaining incrementally smaller values (Fig 1E, F). K-means clustering based on genetic
137 distance showed the strongest support for a single *D. magna* cluster, which is readily
138 observable upon visual inspection of the PCA ordination (Fig 1E).
139
140 In contrast to the host, the majority of genomic variance for *P. ramosa* could be summarized
141 by the first two principal components, which together explained 77 % of total variance (Fig
142 1B). Visual inspection of the first two axes of the PCA ordination showed the *P. ramosa*
143 samples arranged as a triangular cloud with three distinct genomic clusters in the corners,
144 which we have labeled as *Alpha*, *Beta*, and *Gamma* (Fig 1B). Samples infected by multiple *P.*
145 *ramosa* genotypes (Fig 1A) were located between the clusters, primarily along the edges of
146 the triangle. K-means clustering based on genetic distance supported the division of single-
147 infection samples into the *Alpha*, *Beta*, *Gamma* clustering scheme, with the possibility for
148 further subdivision of the *Alpha* lineage into three sub-clusters (Fig 1B, 1C). Admixture
149 analysis of our *P. ramosa* samples showed the strongest support for three clusters, with group
150 membership identical to the *Alpha*, *Beta*, and *Gamma* clustering scheme derived from DAPC
151 analysis (S2 Fig). For the sake of clarity, we hereafter refer to the three *P. ramosa* genomic
152 clusters as "lineages" without reference to specific taxonomic rank.
153
154 The examination of further principal components revealed a small degree of relatively
155 continuous within-lineage variation, potentially reflecting the process of recombination
156 within lineages (Fig 1C). Recombination is not well understood in *P. ramosa,* but we could
157 observe its signature in the form of linkage disequilibrium dissolution with distance across
158 the *P. ramosa* genome (S3 Fig). Admixture analysis indicated strong support for
159 recombination within lineages, but much less among the three primary lineages (S3 Fig).
160 Therefore, the boundaries between sub-clusters should be understood as approximation rather
161 than the clear divisions that we observe between the discrete *Alpha*, *Beta*, and *Gamma*
162 lineages.
163
164 Population structure is an important consideration in GWAS modeling, and is often included
165 as a covariate in the form of principal components or a relatedness matrix (Zhou and
166 Stephens 2012; Chang et al. 2015). This paradigm can be extended to co-GWAS models
167 through inclusion of principal components from both the host and parasite (Ansari et al.
168 2017; Naret et al. 2018). However, several unique features of our system require careful
169 consideration regarding the treatment of population structure in a co-GWAS context. In this

study, the *D. magna* samples originated from a single population absent of geographic subdivision (a small pond). Under this demographic scenario, the major principal components are likely to represent local genomic features rather than genome-wide structure. The inclusion of these PCs as covariates in co-GWAS model might therefore mask highly differentiated loci which are potentially of biological interest from model results. This is even more the case for the parasite, as the strong lineage structure that we found for *P. ramosa* may be tightly linked to (or even derive from) the infection phenotype of the parasite. If so, the top principal components for the parasite would represent a large portion of the biologically relevant aspects of its interaction with the host. Correcting for population structure might therefore mask the biological associations at the core of the interaction.

Co-GWAS

In consideration of these aspects, we have produced three hierarchically nested co-GWAS models that examine host-parasite interlinkage under different assumptions regarding population structure, dominance, and host-specificity. Given the large number of pairwise associations that are tested under each model ($>10^9$) we have avoided setting a specific p-value threshold to assign statistical significance to particular model results. Rather, the strongest peaks from each model were considered as candidate loci for phenotypic confirmation through laboratory-based infection trials.

Our first co-GWAS model contained controls for host and parasite population structure, with the top 10 principal components from the host (22 % total variance) and top 2 principal components from the parasite (82 % total variance) as model covariates. The results from this model showed a single strong peak in the *Daphnia* genome located precisely at the location of a well-characterized locus for *P. ramosa* resistance - the ABC locus (Table 1). The ABC locus is considered a supergene, but we likely observe only the effect of the C locus here because the A locus shows no variation in our study population, and the B locus is nearly fixed for one allele (Ameline et al. 2021). This genomic region showed a strong signal of interlinkage with a tightly clustered set of genes encoding triplets of collagen-like proteins in the *P. ramosa* genome (Fig 2). This small region contains the densest concentration of collagen-like genes in the *P. ramosa* genome, containing PCL triplets [9-10-11], [12-13-14], and [36-37-38] (Fig 2, S4). A single, apparently intragenic, SNP on the right arm of chromosome 5 of the host genome also appeared among the top results for this model (Fig 2, Table 1), but the complete absence of neighboring variants with similar small p-values indicates that this peak is relatively weak candidate (S5 Fig).

Our second co-GWAS model included covariates from the host only, thereby allowing parasite populations structure to be incorporated into the genomic terms of the model. This model showed the same C locus / PCL interlinkage observed in model 1, and a further strong signal of interlinkage precisely at the location of the D locus for *P. ramosa* resistance (Fig 2). The D locus appears to be tightly interlinked with a narrow genomic region containing the PCL gene triplet [21-27-28] (Table 2). We also observed a single, apparently intragenic, host SNP on chromosome 1 among the top results for this model. Like the isolated SNP observed in model 1, the complete absence of neighboring variants with similar p-values suggest that this peak is relatively weak candidate.

Our third co-GWAS model accounted for the fact that our host population is best represented as a single well-mixed population, and further omited all principal components for the host. Model 3 also allowed for dominance effects in the host genotype, which could not be incorporated in previous models due to convergence problems with the larger number of parameters estimated. The strongest signatures of interlinkage in model 3 were again located between previously-characterized *Pasteuria*-resistance loci in the host genome and PCL gene triplets in the parasite genome (Fig 2). The E locus region for *P. ramosa* resistance appeared as one of the largest peaks in model 3, and showed strong interlinkage with PCL triplets [6-7-8], and [36-37-38]. We also observed a strong signal of interlinkage between the D locus for *P. ramosa* resistance and PCL [24-25-26]. In addition to these known *Pasteuria*-resistance loci, we observe a strong signature of interlinkage at two previously-uncharacterized positions in the host genome. The largest of these peaks was found across a broad section of the right arm of chromosome 5, and was associated with PCL triplets [6-7-8], [21-27-28], [24-25-26], and [36-37-38]. The extreme width of this peak is unusual, and perhaps indicates some type of structural polymorphism. The second uncharacterized peak in model 3 occurred between a region on the left arm of host chromosome 7 and PCL triplet [15-16-17]. Neither of these loci have been previously observed to relate to *P. ramosa* resistance in laboratory experiments, but they appear here as strong candidates for further investigation based on the width of the peaks and their highly specific PCL triplet association.

In order to assess the correspondence between host and parasite phenotype, we conducted a series of infection trials using clone lines established from our wild-collected daphnia and a panel of *P. ramosa* isolates. This panel was comprised of five laboratory-cultured *P. ramosa* isolates (C1, C19, P15, P20, and P21), which in various combinations reveal the phenotype of

238 the known pasteuria-resistance loci (Luijckx et al. 2013; Bento et al. 2017; Bento et al. 2020;
239 Ameline et al. 2021). We then applied DAPC clustering to the genomic regions
240 corresponding to each of the resistance loci, and grouped the genomic clusters by resistotype
241 to define functional genotype groups (S6, S7, S8 Tables). For example, the allelic state at the
242 *D. magna* C locus is defined in relation to susceptibility to *P. ramosa* isolates C1 and C19,
243 with resistance being dominant (Bento et al. 2017). Among our samples we observed perfect
244 correspondence between genotype group at the C locus and resistance phenotype, with one
245 group (C') showing complete susceptibility to C1 and C19 and the other (C) showing
246 resistance to C1 and C19 (S6 Table). We observed only partial correspondence between
247 resistotype and genotype for the D and E loci. In both cases, one genotype cluster (D or E)
248 corresponded to exactly one resistance phenotype, but the alternative genotype cluster (D' or
249 E') corresponded to multiple resistance phenotypes (S7, S8 Table). This imperfect
250 correspondence between genotype and phenotype does not appear to be an artifact of our
251 clustering methodology or dominance effects, as apparently identical genotypes were
252 associated with multiple resistance phenotypes. These results indicate that yet unknown loci
253 may interact with the D and E loci to affect *P. ramosa* resistance in a manner which has not
254 been observed during previous laboratory-based experiments.
255
256 *Uncharacterized loci recovered from co-GWAS*
257 Our co-GWAS models recovered a strong signal of interlinkage at two previously
258 uncharacterized loci in the *D. magna* genome. The strongest of these peaks spanned a large
259 (~ 4 Mb) region on the right arm of chromosome 5, while the other spanned ~ 100 Kb on the
260 left arm of chromosome 7. Given the imperfect sorting of D and E locus genotypes into their
261 expected phenotypes, we hypothesized that one or both of these uncharacterized loci could
262 represent an unknown *Pasteuria*-resistance locus that epistatically interacts with either locus.
263 We therefore performed a series of simple GWAS experiments to determine if these loci
264 correspond to different pasteuria resistance phenotypes. We discovered that the candidate
265 locus on the right arm of chromosome 5 is strongly associated with the resistotype RRS
266 relative to the *P. ramosa* isolates C1, C19, and P20 (Fig 3). This is the canonical resistotype
267 used to identify the allelic state at the E locus in laboratory-based experiments (Ameline et al.
268 2021), and provides further evidence that this previously uncharacterized locus may
269 epistatically interact with other loci to affect resistance phenotype. We could not identify a
270 similar correlation between resistance phenotype and the second uncharacterized locus on

chromosome 7, but future experiments are planned to further examine this possibility through an expanded panel of *P. ramosa* isolates.

Genotype – Genotype matrix

Host specificity in the *D. magna* / *P. ramosa* system is hypothesized to arise from relatively simple host-parasite infection matrices, such as a matching allele matrix, but the presence of multiple interacting loci in both host and parasite genomes has the potential to yield considerable complexity. Nonetheless, our functional grouping of hosts according to resistance locus genotypes revealed strong associations with the parasite genomic clusters. Chi-squared tests of genotype contingency tables showed significant associations for the C locus ($X^2 = 16.4$, $p = 2.5^{-3}$), D locus ($X^2 = 32.5$, p-value = 51.5^{-6}), and E locus ($X^2 = 55.2$, $p = 3.0^{-11}$), with deviations from expected values resembling a matching allele model (Fig 4). For example, we observed that *D. magna* carrying the D genotype were most likely to be infected by parasites from the *Gamma* lineage, but apparently immune to infection by *Gamma* when carrying the D' genotype (Fig 4). Similarly, the E genotype appears to confer strong resistance to the *Beta* lineage, while the alternative E' genotype confers strong susceptibility. This specificity remains observable at finer scales, with various sub-clusters of the *Alpha* lineage showing strong differences in ability to infect different C and D locus genotypes.

DISCUSSION

Summary

Our results showed a clear signal of genomic interlinkage between several loci in the *D. magna* genome and a set of genes encoding for collagen-like protein in *P. ramosa.* Our findings were supported by our *P. ramosa* attachment assays, which showed a strong correspondence between phenotype and genotype at the recovered resistance loci. The fact that our co-GWAS model recovered all known resistance loci of *Daphnia* that segregate in our study population stands as a strong biological validation of our results, as experiments describing these loci employed genetic crosses and phenotyping experiments using a laboratory-based panel of hosts and parasites (Metzger et al. 2016; Bento et al. 2017; Bento et al. 2020; Ameline et al. 2021). Our co-GWAS models recovered these loci using only genomic data from wild-collected specimens, with phenotypic data employed only for post-hoc confirmation. While most of these loci in the host were known previously, their interactions with infectivity loci in the parasite were entirely unknown. Our experiments

305 revealed that these are not general resistance and infectivity loci, but rather their effect is
306 contingent on the allelic state at interacting loci in the host or parasite. We also identified a
307 new candidate locus for *Pasteuria*-resistance in the host genome which strongly correlates
308 with resistance to *P. ramosa* isolate P20. A further important finding is that collagen-like
309 gene triplets in the *P. ramosa* genome (but not single collagen-like genes) are key
310 determinants of infectivity. These new candidate loci for infectivity and resistance will be
311 examined through future experiments with genetic crosses and QTL mapping.
312
313 *Co-GWAS model differences*
314 Our co-GWAS modeling revealed that different assumptions regarding population structure,
315 dominance, sample relatedness, and other factors yielded important differences in the results.
316 Therefore, we present not a single model that shows only a snapshot of this parameter space,
317 but present three co-GWAS models which covers the range of variation that we observe
318 under a range of biologically reasonable assumptions. Our first co-GWAS model includes
319 principal components from both the host and parasite as covariates. As parasite lineage
320 structure is strongly correlated with the first two principal components of the *P. ramosa*
321 genome, we primarily observed the effects of within-lineage variation in model 1. This is
322 supported by our host genotype – parasite lineage association tables (Fig 9) which show that
323 the effect of the host's C locus is most apparent across sub-groups within the *Alpha* lineage.
324
325 Our second co-GWAS model includes covariates from the host only, thereby allowing
326 parasite populations structure to be incorporated into the genomic terms of the model. This
327 approach could be considered an analytical mistake, as it could increase the rate of false
328 positives due to non-adaptive demographic effects such as co-dispersal of host and parasite
329 (Naret et al. 2018). However, under such a scenario all lineage specific alleles would show an
330 equally strong association. We do not observe this phenomenon in our results for model 2,
331 rather we observe clearly distinct peaks at specific regions of the host and parasite genome
332 which contain genes previously hypothesized to govern the infection process. Model 2
333 reflects the hypothesis that population structure in *P. ramosa* is a direct result of genetic
334 interactions with its host rather than a confounding factor.
335
336 Our third co-GWAS omitted host and parasite population structure under the assumption that
337 specific combinations of host and parasite genomes occur due to biological rather than
338 demographic reasons. This assumption is reasonable given that all samples were collected

339 from one population during one season, but would be advised against when populations are
340 comprised of discrete demes or drawn from a wide geographic range. In this model we
341 observed strong peaks of interlinkage between *Pasteuria*-resistance loci and PCL gene
342 triplets, and identified two additional candidate loci in the host genome showing similar
343 signals of interlinkage. Subsequent laboratory experiments showed one of these loci to be
344 highly correlated with resistance to *P. ramosa* isolate P20. Although each of these models
345 vary to some degree in specific details, we observe a consistent pattern of *Pasteuria*-
346 resistance genes in the host showing a highly specific association with collagen-like genes in
347 the *P. ramosa* genome that are arranged in a conserved triplet configuration.
348
349 *Signatures of balancing selection at interlinked loci*
350 Negative-frequency dependent selection is a form of balancing selection that is expected to
351 yield a genomic signature of locally elevated nucleotide diversity at the sites under selection
352 (Ebert and Fields 2020). This pattern arises because common host variants are selected
353 against by parasite variants that specialize on these host variants. A previous study of the
354 genetic basis of *P. ramosa* infection found usually high levels of genetic diversity at a
355 number of PCL genes, which the authors interpreted as evidence of balancing selection at
356 those loci (Andras et al. 2018). Samples for this previous study were drawn from a wide
357 geographic range that spanned across the Western Palaearctic. Despite the fact that our study
358 samples were drawn from a single population, we observed peaks in nucleotide diversity at
359 the same genomic positions (Fig 2). In respect to the *D. magna* host, we observed the highest
360 levels of nucleotide diversity at the position of the ABC locus, which is known to contain
361 major structural variants of sufficient divergence to limit recombination (Bourgeois et al.
362 2021). Thus, we observe evidence for balancing selection in terms of nucleotide diversity as
363 well as genomic structural diversity among both the host and parasite dimensions of this
364 system.
365
366 *Molecular mechanisms of infection and resistance*
367 Pasteuria collagen-like proteins (PCLs) belong to a larger class of collagen-like proteins
368 which are known to occur in at least 14 bacterial phyla (Qiu et al. 2021). For example, many
369 species of pathogenic *Streptococcus* contain collagen like proteins (Xu et al. 2002; Karlström
370 et al. 2004; Paterson et al. 2008; Lukomski et al. 2017), as does the highly virulent *Bacillus*
371 *anthracis* (Henriques and Moran 2007). Collagen-like proteins have been directly linked to
372 infectivity in the gut pathogen *Clostridium difficile* (Strong et al. 2014), drug resistant strains

373 of the human and livestock parasites *Burkholderia* sp. (Bachert et al. 2015), and the causative
374 agent of Legionnaire's disease, *Legionella pneumophila* (Vandersmissen et al. 2010).
375 Collagen-like proteins are hypothesized to mediate the infection process through the
376 mechanism of attachment to host tissues (Qiu et al. 2021). Compared to other bacterial
377 species, *P. ramosa* has an unusually large number of collagen-like proteins, with nearly 50
378 genes identified thus far (McElroy et al. 2011). A previous study found that variation at one
379 of these PCLs correlated perfectly with the ability of certain *P. ramosa* isolates to infect a
380 specific *D. magna* genotype (PCL 7 of the triplet [6-7-8]) (Andras et al. 2020). Given that
381 *Pasteuria* spores are coated with a thick layer of collagen-like fibers (Davies 2009),
382 recognition and attachment to host tissue is a highly plausible role for these proteins. The
383 enormous repertoire of collagen-like proteins in *P. ramosa* could be explained as a long-term
384 effect of balancing selection due to a tight coevolutionary history with its host. A noteworthy
385 finding is that the collagen-like genes revealed by our co-GWAS models all appeared in a
386 specific triplet formation. The *P. ramosa* genome contains many isolated collagen genes and
387 loose aggregations of collagen genes, but none of these showed strong association with the
388 host. Thus, with regard to infection phenotype, we observed that not only the nature of the
389 gene is important, but also the spatial orientation of the gene in the genome. The function of
390 the collagen-like gene triplets, as opposed to single genes is not known.
391
392 The molecular mechanisms which determine *D. magna* resistance to specific *P. ramosa*
393 genotypes are not yet resolved, although a promising hypothesis concerns the post-
394 translational modification of cuticle proteins via the attachment of branching sugars (*i.e.,*
395 glycosylation). Previous studies have found major differences between susceptible and
396 resistant clones at the ABC locus in regards to genes encoding fucosyltransferases,
397 glucosyltransferases, galactosyltrasferases – all of which are involved in the post-translational
398 attachment of sugars to proteins (Bento et al. 2017). The E locus, which appears have an
399 independent evolutionary origin from the ABC locus, seems to be enriched for
400 galactosyltransferase and glucosyltransferase genes as well (Ameline et al. 2021). In an
401 interesting parallel, it has been proposed that glycosylation of amino acid residues at the C-
402 terminal domain also plays an important role in attachment specificity of bacterial collagen-
403 like proteins (Strong et al. 2014; Qiu et al. 2021). There is some support for this hypothesis
404 from a previous study of *P. ramosa* infectiousness, which found that a predicted N-
405 glycosylation site in the C-terminal domain of PCL 7 correlated perfectly with infection
406 phenotype (Andras et al. 2020).

In conclusion, understanding the genomic basis of infectious disease is a fundamental objective in the study of coevolution. Our results show a clear signal of interspecies linkage disequilibrium across multiple sets of interacting loci in a naturally coevolving system, thus confirming the most fundamental assumption of host-parasite coevolution. Previous co-genomic association studies have identified variants of clinical significance for HIV infections (Bartha et al. 2013), Hepatitis C (Ansari et al. 2017), pneumococcal meningitis (*16*), and malaria (Band et al. 2021) through a combination of SNP arrays and patient clinical data. Our approach moves a step further by examining the complete genomes of naturally infected hosts and parasites from a host population with a known history of parasite epidemics. Our central finding is an underlying host-parasite genetic interaction matrix, as such matrices are at the core of coevolutionary theory, and their topology determines the tempo and mode of coevolution. Our matrix confirms the importance of multiple loci in both antagonists, acting together within and across genomes, with epistasis playing a decisive role for the host and the parasite. We demonstrate that the co-genomic framework can be successfully applied to non-model systems, and given sufficiently strong host-parasite specificity, resolution of complex traits is possible in a candidate-free manner. In the future, this method will be useful to elucidate other types of pairwise interactions, such as commensalism, mutualism, predation, and assortative mating in the context of sexual selection.

MATERIALS AND METHODS

Field Collections

We collected *D. magna* samples from Lake Aegelsee (47.558048, 8.861063) via horizontal net tows at the deepest place during a strong *P. ramosa* epidemic in summer 2019. These wild-collected *Daphnia* were monitored in the laboratory for signs of *P. ramosa* infection and the appearance of offspring. Although *P. ramosa* typically castrates the host, clonal reproduction is still possible in the early stages of infection. As *P. ramosa* is exclusively transmitted through the environment, the laboratory lines were quickly separated from the mother and remained uninfected throughout the experiment. After verifying *P. ramosa* infection status and establishing healthy clonal lines, we preserved the original wild-collected individuals in 90 % EtOH at -20 °C for later DNA extraction.

441 *DNA extraction and Sequencing*
442 Whole-organism DNA extractions from *Daphnia* may contain a high proportion of non-target
443 DNA from undigested food, gut microbiota, and bacteria microfilm on the cuticle. To reduce
444 contaminants, we cleared the gut by feeding the daphnia with a suspension of sterile beads
445 (Sephadex G-25, cross-linked dextran gel) three times per day for at least 24 hours. After the
446 gut was visibly cleared of food, we monitored containers with individual daphnia every 2-4
447 hours for the presence of a molted cuticle. After molting, *Daphnia* were immediately
448 preserved in 90 % EtOH at -20 °C for later DNA extraction, sharply reducing the amount of
449 bacterial contamination in the preserved sample. Preliminary experiments showed this
450 protocol to be effective in limiting the amount of non-target DNA in preserved samples to
451 less than 5 %.
452
453 We extracted DNA from infected *Daphnia* using a modified protocol for a Gentra PureGene
454 column-based extraction kit. Briefly, we homogenized whole *Daphnia* with a pestle and
455 incubated the samples overnight at 55 °C in a solution of lysis buffer and Proteinase K. The
456 samples were then incubated for 30 minutes at 37 °C with RNase A, treated with a protein
457 precipitation solution, and centrifuged to remove precipitates. The supernatant was then
458 subject to a chilled isopropanol precipitation with glycogen added as a DNA carrier. The
459 DNA pellet was then resuspended in a buffer solution and stored at -20 °C. Our complete
460 extraction protocol is available on the GitHub repository for this manuscript
461 (https://github.com/edexter/Interlink).
462
463 We then prepared genomic libraries for short-read sequencing using a NEBNext Ultra II
464 DNA Library Prep Kit for Illumina. The libraries were sequenced across multiple runs of a
465 Novaseq 6000 to generate paired-end 150 bp reads, aiming for at least 30X coverage across
466 host and parasite genomes. Samples were sequenced with individual barcodes for subsequent
467 demultiplexing and inspection for sequencer batch and lane effects. Genomic library
468 preparation and Illumina sequencing were performed at the Basel Genomics Facility, Basel,
469 Switzerland.
470
471 *Host phenotyping*
472 The most important step of the *P. ramosa / D. magna* infection process is attachment of *P.*
473 *ramosa* spores to the any of several sites on the cuticle, and may occur through the uptake of
474 *P. ramosa* spores during feeding (Duneau et al. 2011; Fredericksen et al. 2021). Attachment

475 is a necessary requirement for succeful infection and variation in attachment explain the vast
476 majority of overall infection success (Hall et al. 2019, Molecular Ecology). Attachment
477 specificity between host and parasite genotype combinations can be determined via a
478 relatively simple attachment test using fluorescently labeled *Pasteuria* spores. Experimental
479 studies based upon this attachment test have found a nearly perfect correlation between
480 attachment and infection outcome, with little to no effect of environmental or host condition
481 (Duneau et al. 2011).
482
483 We collected resistance phenotype data of the host (aka. "resistotype") for a subset of 107
484 genotypes of our sequenced *D. magna* clones, with each clone tested in triplicate. *Daphnia*
485 resistotype was assessed through gut attachment tests with fluorescently labeled *Pasteuria*
486 spores using a standardized panel of five *Pasteuria* isolates: C1, C19, P15, P20, and P21
487 (Duneau et al. 2011; Bento et al. 2017; Bento et al. 2020; Ameline et al. 2021). Resistance
488 (R) or susceptibility (S) to each isolate is denoted in a five-letter resistotype code with each
489 position corresponding to a single isolate. For example, the resistotype [SRRRR] would
490 indicate susceptibility only to C1, while the resistotype [SSSRR] would indicate
491 susceptibility to C1, C19, and P15. These attachment data were used for *post hoc* validation
492 of our co-GWAS results through comparison of genotypes at interlinked loci and resistance
493 phenotypes.
494
495 *Bioinformatics*
496 De-multiplexed sequence-read pairs were trimmed to remove low-quality reads and adaptor
497 sequences using Trimmomatic version 0.39 (Bolger et al. 2014). Duplicate reads were then
498 removed using Picard version 2.9.2. The trimmed reads were mapped against host and
499 parasite reference genomes using BWA-MEM version 0.7.15 operating under default
500 parameters (Li 2013). Local indel realignment was then performed on the mapped reads using
501 GATK version 4.2.0.0 (Auwera and O'Connor 2020). Reads mapping to both host and
502 parasite genome were flagged for downstream filtering. We jointly called genomic variants
503 (SNPs, indels, and mixed-type variants) across all samples using the GATK haplotype caller
504 following the best practices pipeline for diploid samples (Auwera and O'Connor 2020).
505 Although *P. ramosa* is haploid, this approach enabled us to differentiate single genotype
506 infections from multiple infections, and to identify major and minor alleles within each pool.
507 *In silico* experiments showed little change in the major/minor alleles identified when calling
508 *P. ramosa* variants under different ploidy models.

509 The host and parasite VCF files were filtered to remove low-quality bases, poor alignment
510 scores, or extreme strand biases. Variants were then filtered to exclude sites with excessive
511 depth (> 2x mean depth), or excessive instances of dual mapping (reads mapping to both
512 genomes) using a custom R script. Based upon visual inspection of dual mapping regions,
513 low-complexity sequences rather than assembly errors or gene homology appear to be the
514 most likely cause of dual-mapping. The VCF files were then filtered to remove multi-allelic
515 variants and converted to PED files for input to Plink for co-GWAS analysis. Summary
516 statistics related to read mapping were calculated with SAMtools version 1.10 (Li et al.
517 2009). We calculated nucleotide diversity (π) for host and parasite genomes with VCFtools
518 version 0.1.16 (Danecek et al. 2011). For the host, we calculated π across 100 Kb sliding
519 windows with a 1 Kb offset. Across the much smaller parasite genome, we calculated π
520 across 500 bp sliding windows with a 250 bp offset. Precise filter cutoff values and other
521 parameters can be found along with our complete code at the associated Github repository.
522
523 *Co-GWAS analysis*
524 We employed a co-genome wide sequencing association study (co-GWAS) framework to
525 search for interlinked genomic regions of the host and parasite. This approach is based on the
526 regression of host genotypes against parasite genotypes for every pairwise combination of
527 loci. Similar methods of genome-to-genome association testing have been previously
528 employed to find clinically relevant variants in HIV (Bartha et al. 2013), Hepatitis C (Ansari
529 et al. 2017), *Streptococcus pneumoniae* (Lees et al. 2019), and *Plasmodium falciparum* (Band
530 et al. 2021) using human SNP array data. We prepared our genomic data for co-GWAS by
531 removing variants with a minor allele frequency less than 0.05 or missingness greater than
532 0.20. Pairwise relatedness was calculated for the host using the KING kinship estimator, and
533 samples showing relatedness greater than 0.35 (possible clone mates) were thinned to remove
534 redundant samples. To reduce the probability of false positives due to population structure,
535 we conducted PCA on subsets of unlinked loci for both host and parasite. The top principal
536 components were then included as covariates in our co-GWAS models. We used Plink
537 version 2.00a3LM for PCA, frequency-based filtering, and co-GWAS analysis (Chang et al.
538 2015).
539
540 Our GWAS analysis was structured as a logistic regression model of the form:
541
542 $$Y_{ij} \sim \mathbf{X}k + \mathbf{Z} + e$$

543 Where Y is a binary variable [0,1] indicating the presence or absence of allele i at locus j in
544 the parasite genome. The variable X represents the genotype at locus k in the host genome
545 expressed as the major allele dosage [0,1,2], a dominant reference allele model [1,1,0], or a
546 dominant alternate allele model [0,1,1]. These different coding schemes reflect the fact that
547 host samples contain a single diploid individual while their infecting parasites may contain a
548 pool of haploid individuals (i.e., multiple infections). Note that although dominance models
549 are more biological plausible than a simple genotypic dosage model, they are rarely
550 employed in the context of GWAS due to statistical difficulties in detecting non-additivity
551 (i.e., too few homozygous-minor genotypes) (Chang et al. 2015; Marees et al. 2018). The
552 term Z [0-1] represents a matrix of the top n principal components and controls for population
553 structure among samples, with permutations of our primary co-GWAS model containing
554 different subsets of principal components. In order to reduce the computational burden of
555 weak or non-significant associations, test results with p-values greater than 10^{-4} were not
556 stored for downstream analysis or plotting. This cutoff was derived from preliminary co-
557 GWAS trials which showed that p-values below this threshold were uniformly distributed
558 across the *Daphnia* genome and generally uninformative.
559
560 Because each variant position in the host genome is tested against each variant position in the
561 parasite genome, co-GWAS analysis can generate a considerable number of false positives
562 purely through stochastic processes. Rather than applying a simple threshold to denote
563 statistical significance, we utilized information about the magnitude and breadth of each peak
564 to create a list of candidate loci for further examination. Specifically, biologically relevant
565 associations are expected to yield a broad peak of elevated p-values, including nearby
566 flanking regions which rapidly decay with distance from the site of selection. This pattern is a
567 characteristic signature of linkage disequilibrium decay due to recombination and provides a
568 useful biological validation to our statistical results. In contrast, we expect that associations
569 due to mapping errors or stochastic sampling effects will result in isolated peaks comprised
570 by one or very few variants.
571
572 *Host and parasite genotype clustering*
573 We extracted the sequence data from the genomic region corresponding to each of the known
574 host resistance loci and applied a discriminant analysis of principal components (DAPC)
575 function to aggregate individual genotypes into clusters of broadly similar genotypes
576 (Jombart et al. 2010). Genotype clusters were then grouped by resistance phenotype to define

577 functional groups. Each host was assigned a set of three resistance locus codes denoting the
578 functional group of the C, D, and E locus, each with two functional groups (e.g., C vs C'). We
579 applied the same DAPC method to our parasite samples, using the entire parasite genome
580 rather than individual loci. DAPC and PCA were performed using the Adegenet version 2.1.3
581 package for R version 4.0.4 (Jombart 2008). All statistical analyses were performed using R
582 unless specifically mentioned otherwise, and all figures were produced using the ggplot2
583 (version 3.3.3) package (R Core Team 2016; Wickham et al. 2018). DAPC clustering results
584 were compared against genomic clusters generated via the K-means cross-validation
585 algorithm implemented in Admixture version 1.3.0 (Alexander et al. 2009).

Acknowledgments

We would like to thank the Ebert lab members for valuable discussions and suggestions.

Funding

This work was supported by the European Commission Marie Curie Postdoctoral Fellowship grant #835379 (ED, DE) and the Swiss National Science Foundation (SNSF) grant #310030_188887 (DE)

Author contributions:

Conceptualization: ED, DE, PDF
Methodology: ED, DE, PDF
Investigation: ED, DE
Visualization: ED
Funding acquisition: ED, DE
Project administration: ED, DE
Supervision: DE
Writing – original draft: ED, DE
Writing – review & editing: ED, DE, PDF

Competing interests

Authors declare that they have no competing interests

Data and materials availability

The complete annotated code necessary to reproduce all analyses and figures presented here is available on GitHub (https://github.com/edexter/Interlink). The raw genomic data pertaining to this study are available from the NCBI Short Read Archive (SRA) (accession number in progress).

614 **REFERENCES**

615 Agrawal A, Lively CM. 2002. Infection genetics: gene-for-gene versus matching alleles
616 models and all points in between. *Evol. Ecol. Res.* 4:91–107.
617 Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in
618 unrelated individuals. *Genome Res.* 19:1655–1664.
619 Ameline C, Bourgeois Y, Vögtli F, Savola E, Andras J, Engelstädter J, Ebert D. 2021. A two-
620 locus system with strong epistasis underlies rapid parasite-mediated evolution of host
621 resistance. *Mol. Biol. Evol.* 38:1512–1528.
622 Andras JP, Fields PD, Du Pasquier L, Fredericksen M, Ebert D. 2020. Genome-wide
623 association analysis identifies a genetic basis of infectivity in a model bacterial
624 pathogen. *Mol. Biol. Evol.* 37:3439–3452.
625 Andras JP, Fields PD, Ebert D. 2018. Spatial population genetic structure of a bacterial
626 parasite in close coevolution with its host. *Mol. Ecol.* 27:1371–1384.
627 Ansari MA, Pedergnana V, Ip CLC, Magri A, Von Delft A, Bonsall D, Chaturvedi N, Bartha
628 I, Smith D, Nicholson G, et al. 2017. Genome-to-genome analysis highlights the
629 effect of the human innate and adaptive immune systems on the hepatitis C virus. *Nat.*
630 *Genet.* 49:666–673.
631 Auwera GAV der, O'Connor BD. 2020. Genomics in the Cloud: Using Docker, GATK, and
632 WDL in Terra. O'Reilly Media, Inc.
633 Bachert BA, Choi SJ, Snyder AK, Rio RVM, Durney BC, Holland LA, Amemiya K, Welkos
634 SL, Bozue JA, Cote CK, et al. 2015. A unique set of the Burkholderia collagen-like
635 proteins provides insight into pathogenesis, genome evolution and niche adaptation,
636 and infection detection. *PLOS ONE* 10:e0137578.
637 Band G, Leffler EM, Jallow M, Sisay-Joof F, Ndila CM, Macharia AW, Hubbart C, Jeffreys
638 AE, Rowlands K, Nguyen T, et al. 2021. Malaria protection due to sickle
639 haemoglobin depends on parasite genotype. *Nature*:1–9.
640 Bartha I, Carlson JM, Brumme CJ, McLaren PJ, Brumme ZL, John M, Haas DW, Martinez-
641 Picado J, Dalmau J, López-Galíndez C, et al. 2013. A genome-to-genome analysis of
642 associations between human genetic variation, HIV-1 sequence diversity, and viral
643 control. *eLife* 2:e01123.
644 Bento G, Fields PD, Duneau D, Ebert D. 2020. An alternative route of bacterial infection
645 associated with a novel resistance locus in the Daphnia–Pasteuria host–parasite
646 system. *Heredity* 125:173–183.

647 Bento G, Routtu J, Fields PD, Bourgeois Y, Pasquier LD, Ebert D. 2017. The genetic basis of
648 resistance and matching-allele interactions of a host-parasite system: The Daphnia
649 magna-Pasteuria ramosa model. *PLOS Genet.* 13:e1006596.
650 Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina
651 sequence data. *Bioinformatics* 30:2114–2120.
652 Bourgeois Y, Fields PD, Bento G, Ebert D. 2021. Balancing Selection for Pathogen
653 Resistance Reveals an Intercontinental Signature of Red Queen Coevolution. *Mol.*
654 *Biol. Evol.* 38:4918–4933.
655 Carlson JM, Brumme ZL, Rousseau CM, Brumme CJ, Matthews P, Kadie C, Mullins JI,
656 Walker BD, Harrigan PR, Goulder PJR, et al. 2008. Phylogenetic dependency
657 networks: Inferring patterns of CTL escape and codon covariation in HIV-1 Gag.
658 *PLoS Comput. Biol.* 4.
659 Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. 2015. Second-generation
660 PLINK: rising to the challenge of larger and richer datasets. *GigaScience* 4:s13742-
661 015-0047–0048.
662 Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter
663 G, Marth GT, Sherry ST, et al. 2011. The variant call format and VCFtools.
664 *Bioinformatics* 27:2156–2158.
665 Davies KG. 2009. Chapter 9 Understanding the Interaction Between an Obligate
666 Hyperparasitic Bacterium, Pasteuria penetrans and its Obligate Plant-Parasitic
667 Nematode Host, Meloidogyne spp. In: Advances in Parasitology. Vol. 68. Natural
668 History of Host-Parasite Interactions. Academic Press. p. 211–245. Available from:
669 https://www.sciencedirect.com/science/article/pii/S0065308X0800609X
670 Duneau D, Luijckx P, Ben-Ami F, Laforsch C, Ebert D. 2011. Resolving the infection
671 process reveals striking differences in the contribution of environment, genetics and
672 phylogeny to host-parasite interactions. *BMC Biol.* 9:11.
673 Ebert D. 2005. Ecology, epidemiology, and evolution of parasitism in daphnia. Bethesda,
674 Md.: National Library of Medicine (US), National Center for Biotechnology
675 Information. Available from:
676 http://www.ncbi.nlm.nih.gov/books/bv.fcgi?call=bv.View..ShowTOC&rid=daph.TO
677 C&depth=2
678 Ebert D, Fields PD. 2020. Host–parasite co-evolution and its genomic signature. *Nat. Rev.*
679 *Genet.* 21:754–768.

680 Ebert D, Rainey P, Embley TM, Scholz D. 1996. Development, life cycle, ultrastructure and
681 phylogenetic position of Pasteuria ramosa Metchnikoff 1888: rediscovery of an
682 obligate endoparasite of Daphnia magna Straus. *Philos. Trans. R. Soc. Lond. B. Biol.*
683 *Sci.* 351:1689–1701.
684 Fredericksen M, Ameline C, Krebs M, Hüssy B, Fields PD, Andras JP, Ebert D. 2021.
685 Infection phenotypes of a coevolving parasite are highly diverse, structured, and
686 specific. *Evolution* 75:2540–2554.
687 Henriques AO, Moran CP Jr. 2007. Structure, assembly, and function of the spore surface
688 layers. *Annu. Rev. Microbiol.* 61:555–588.
689 Jombart T. 2008. adegenet: a R package for the multivariate analysis of genetic markers.
690 *Bioinformatics* 24:1403–1405.
691 Jombart T, Devillard S, Balloux F. 2010. Discriminant analysis of principal components: a
692 new method for the analysis of genetically structured populations. *BMC Genet.* 11:94.
693 Karlström Å, Jacobsson K, Flock M, Flock J-I, Guss B. 2004. Identification of a novel
694 collagen-like protein, SclC, in Streptococcus equi using signal sequence phage
695 display. *Vet. Microbiol.* 104:179–188.
696 Lees JA, Ferwerda B, Kremer PHC, Wheeler NE, Serón MV, Croucher NJ, Gladstone RA,
697 Bootsma HJ, Rots NY, Wijmega-Monsuur AJ, et al. 2019. Joint sequencing of human
698 and pathogen genomes reveals the genetics of pneumococcal meningitis. *Nat.*
699 *Commun.* 10:1–14.
700 Legendre P, Desdevises Y, Bazin E. 2002. A statistical test for host–parasite coevolution.
701 *Syst. Biol.* 51:217–234.
702 Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-
703 MEM. *ArXiv13033997 Q-Bio*.
704 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin
705 R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence
706 Alignment/Map format and SAMtools. *Bioinformatics* 25:2078–2079.
707 Luijckx P, Fienberg H, Duneau D, Ebert D. 2013. A matching-allele model explains host
708 resistance to parasites. *Curr. Biol.* 23:1085–1088.
709 Lukomski S, Bachert BA, Squeglia F, Berisio R. 2017. Collagen-like proteins of pathogenic
710 streptococci. *Mol. Microbiol.* 103:919–930.
711 Marees AT, de Kluiver H, Stringer S, Vorspan F, Curis E, Marie-Claire C, Derks EM. 2018.
712 A tutorial on conducting genome-wide association studies: Quality control and

statistical analysis. *Int. J. Methods Psychiatr. Res.* [Internet] 27. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6001694/

Märkle H, John S, Cornille A, Fields PD, Tellier A. 2021. Novel genomic approaches to study antagonistic coevolution between hosts and parasites. *Mol. Ecol.* 30:3660–3676.

McElroy K, Mouton L, Du Pasquier L, Qi W, Ebert D. 2011. Characterisation of a large family of polymorphic collagen-like proteins in the endospore-forming bacterium Pasteuria ramosa. *Res. Microbiol.* 162:701–714.

Metzger CMJA, Pepijn L, Gilberto B, Mahendra M, Ebert D. 2016. The Red Queen lives: Epistasis between linked resistance loci. *Evolution* 70:480–487.

Morand S. 2015. (macro-) Evolutionary ecology of parasite diversity: From determinants of parasite species richness to host diversification. *Int. J. Parasitol. Parasites Wildl.* 4:80–87.

Naret O, Chaturvedi N, Bartha I, Hammer C, Fellay J, The Swiss HIV Cohort Study (SHCS). 2018. Correcting for population stratification reduces false positive and false negative results in joint analyses of host and pathogen genomes. *Front. Genet.* 9.

Paterson GK, Nieminen L, Jefferies JMC, Mitchell TJ. 2008. PclA, a pneumococcal collagen-like protein with selected strain distribution, contributes to adherence and invasion of host cells. *FEMS Microbiol. Lett.* 285:170–176.

Poulin R. 2011. Evolutionary ecology of parasites. 2nd ed. Princeton University Press

Qiu Y, Zhai C, Chen L, Liu X, Yeo J. 2021. Current insights on the diverse structures and functions in bacterial collagen-like proteins. *ACS Biomater. Sci. Eng.*

R Core Team. 2016. R: A language and environment for statistical computing. Available from: https://www.R-project.org/

Schmid-Hempel P. 2021. Evolutionary parasitology: the integrated study of infections, immunology, ecology, and genetics. 2nd ed. Oxford University Press

Strong PCR, Fulton KM, Aubry A, Foote S, Twine SM, Logan SM. 2014. Identification and characterization of glycoproteins on the spore surface of Clostridium difficile. *J. Bacteriol.* 196:2627–2637.

Vandersmissen L, De Buck E, Saels V, Coil DA, Anné J. 2010. A Legionella pneumophila collagen-like protein encoded by a gene with a variable number of tandem repeats is involved in the adherence and invasion of host cells. *FEMS Microbiol. Lett.* 306:168–176.

Wang M, Roux F, Bartoli C, Huard-Chauveau C, Meyer C, Lee H, Roby D, McPeek MS, Bergelson J. 2018. Two-way mixed-effects methods for joint association analysis using both host and pathogen genomes. *Proc. Natl. Acad. Sci.*:201710980.

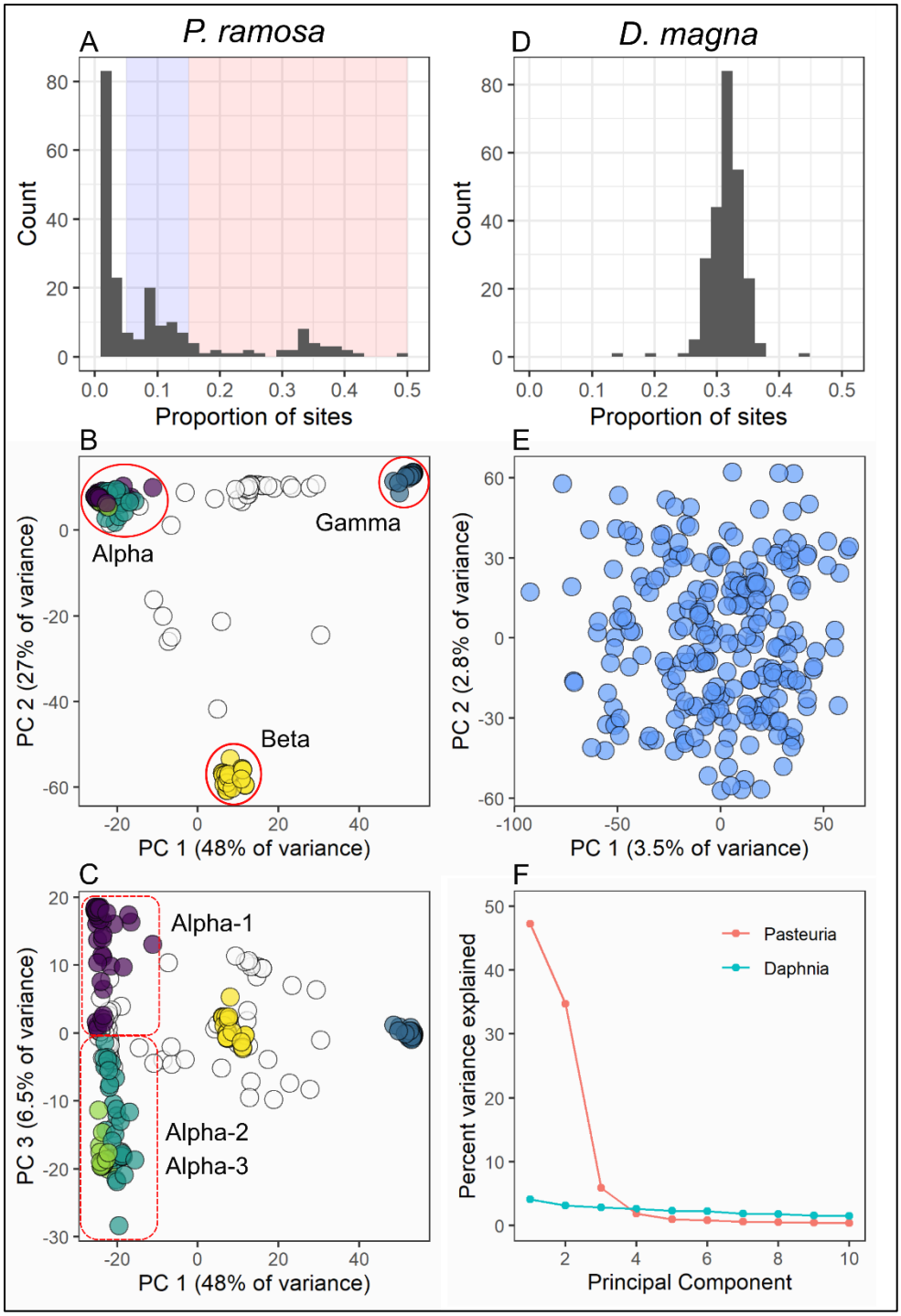
Wickham H, Chang W, Henry L, Pedersen TL, Takahashi K, Wilke C, Woo K, RStudio. 2018. ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics. Available from: https://CRAN.R-project.org/package=ggplot2

Xu Y, Keene DR, Bujnicki JM, Höök M, Lukomski S. 2002. Streptococcal Scl1 and Scl2 proteins form collagen-like triple helices *. *J. Biol. Chem.* 277:27312–27318.

Zhou X, Stephens M. 2012. Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* 44:821–824.

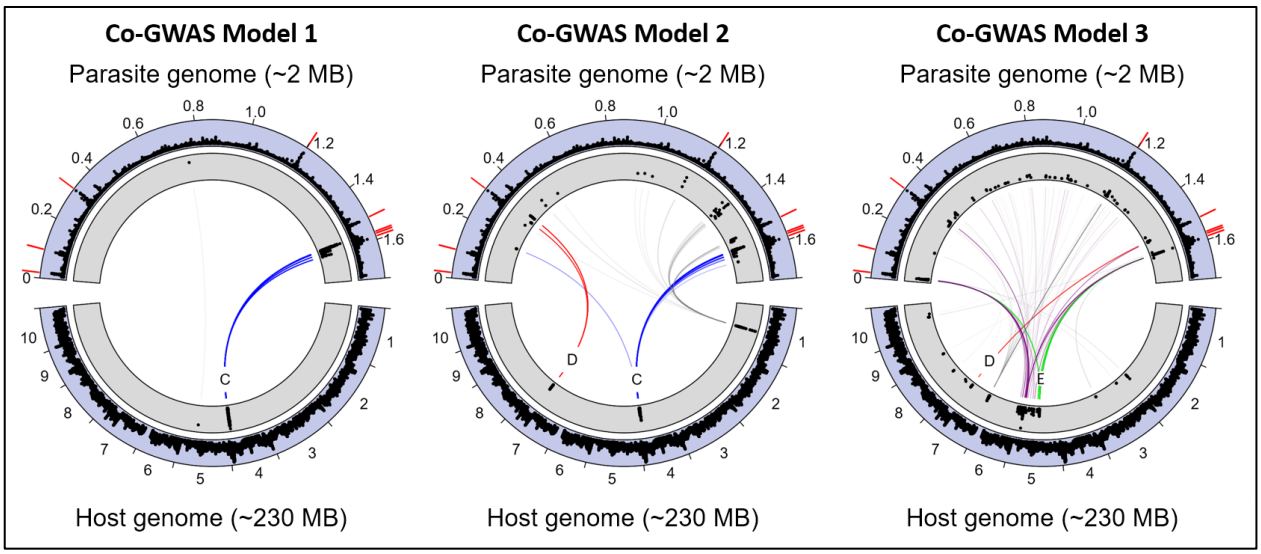
FIGURES

Figure 1. Genomic diversity and population structure of samples used in this study. **(A)** Per sample levels of genomic diversity for *P. ramosa,* expressed as the proportion of population-wide variant sites showing within sample polymorphism. Single genotype infections of the haploid *P. ramosa* are expected to show values near zero, while larger values indicate multiple infections. Values > 0.15 represent infections from multiple distinct pasteuria lineages (shaded in red), with lesser values indicating multiple infections of closely related genotypes (blue). **(B)** PCA ordination of parasite genomes (N = 248) recovered from infected hosts. Parasites cluster into three primary genomic groups (*Alpha*, *Beta*, *Gamma*), with membership derived by DAPC clustering of whole-genome data. Open circles represent multiple infections containing mixtures of genotypes from different clusters (red shaded area on panel A). **(C)** The same PCA ordination rotated to show PCs 1 and 3. The *Alpha* cluster shows relatively continuous variation across the PC 3, which DAPC clustering hierarchically sub-divides into *Alpha-1* and *Alpha-2* / *Alpha-3.* Note the relatively modest contribution to total genomic variation by PC 3 relative to the first two PCs. **(D)** Per sample levels of genomic diversity for *D. magna*, where each sample represents a single diploid individual**.** **(E)** PCA ordination of D*. magna* samples (N = 248) plotted across the first two principal components. Note the low percentage of variance explained by the primary axes and the lack of overall structure. DAPC clustering for *D. magna* best supports a single cluster. **(F)** PCA scree plots for *P. ramosa* and *D. magna.*

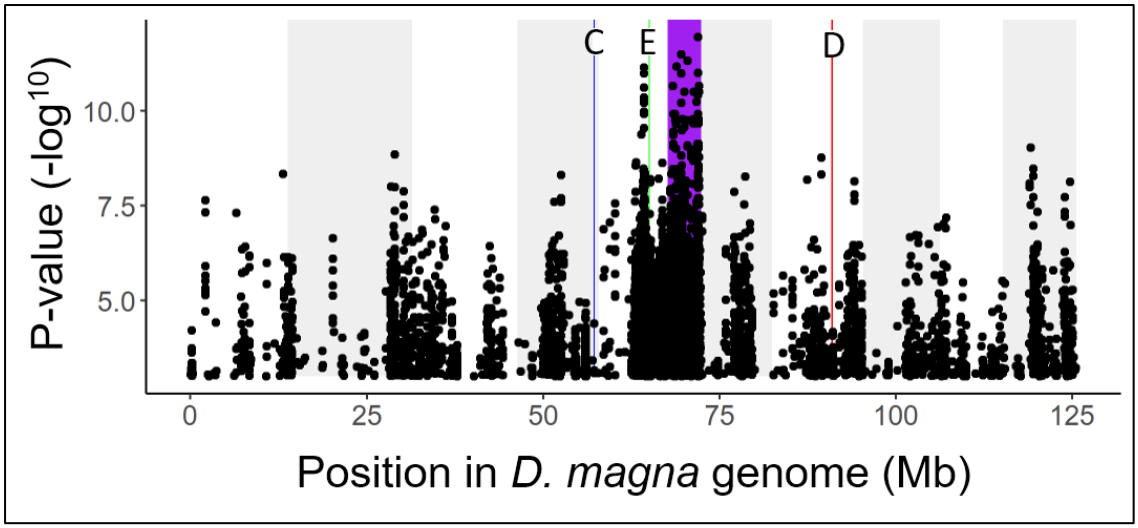

A
P. ramosa
Count
Proportion of sites
D
D. magna
Count
Proportion of sites
B
Alpha
Gamma
Beta
PC 2 (27% of variance)
PC 1 (48% of variance)
E
PC 2 (2.8% of variance)
PC 1 (3.5% of variance)
C
Alpha-1
Alpha-2
Alpha-3
PC 3 (6.5% of variance)
PC 1 (48% of variance)
F
Pasteuria
Daphnia
Percent variance explained
Principal Component

776

Figure 2. The 500 strongest associations for each co-GWAS model between variants in the host genome (bottom ring segments) and the parasite genome (top segments). Inner ring segments (gray background) indicate the scaled genomic position (x-axis) and strength (y-axis) of the association, expressed as the p-value of the co-GWAS score on a -$\log_{10}$ scale. Outer ring segments (with blue background) indicate nucleotide diversity (π), scaled from 0 to 1. Darker pathways indicate a greater number of associations, with known resistance loci of the host being shown in blue, red, and green (C, D, and E locus, respectively). Outer red bars on the *P. ramosa* genome indicate the position of PCL gene triplets; the outer ring of the host genome displays chromosome numbers. Model 1 shows a strong association between the host C locus and a dense cluster of PCL gene triplets in the parasite genome. Model 2 shows the C locus (as in model 1), but also shows a strong association between the host D locus and a different PCL gene triplet in the parasite genome. In model 3 the strongest associations occur on two regions of chromosome 5: One region coincides with the previously described E locus for pasteuria resistance, the other corresponds to a previously undescribed candidate locus for pasteuria resistance which may interact with the E locus (see main text). Lines show in purple pertain to a new candidate locus for *Pasteuria* resistance.



793 **Figure 3.** Single phenotype GWAS for resistance to the *P. ramosa* isolate P20. Resistance or
794 susceptibility to this isolate is diagnostic of the allelic state at the E locus. The positions of
795 the C locus (blue), D locus (red), E locus (green) and new candidate locus for *Pasteuria*
796 resistance (purple) are indicated by vertical bars. Our results show a strong peak at the
797 position of the E locus (green) as expected from previous studies. However, we observe a
798 much stronger peak on the right arm of chromosome 5 (purple) that coincides with the
799 primary peak from our third co-GWAS model. This result provides further evidence of a
800 uncharacterized pasteuria resistance locus at this position, which may epistatically interact
801 with the E locus.



802

Figure 4. Contingency tables of genotype clusters at host resistance loci (derived via DAPC) versus *P. ramosa* genomic cluster. Cells are colored according to the magnitude and direction of deviation from the expected number of observations (Pearson residuals), with host and parasite combinations occurring more often than expected shown in red, and less often in blue. Host genotype clusters are characterized as resistant or susceptible relative to their diagnostic *P. ramosa* isolates – none are universally susceptible or resistant. Residuals for host-parasite combinations were calculated from a chi-squared contingency table test ($p < 0.001$ for all matrices, see text for full result). Multiple infections (open circles on panel B) are not included here.

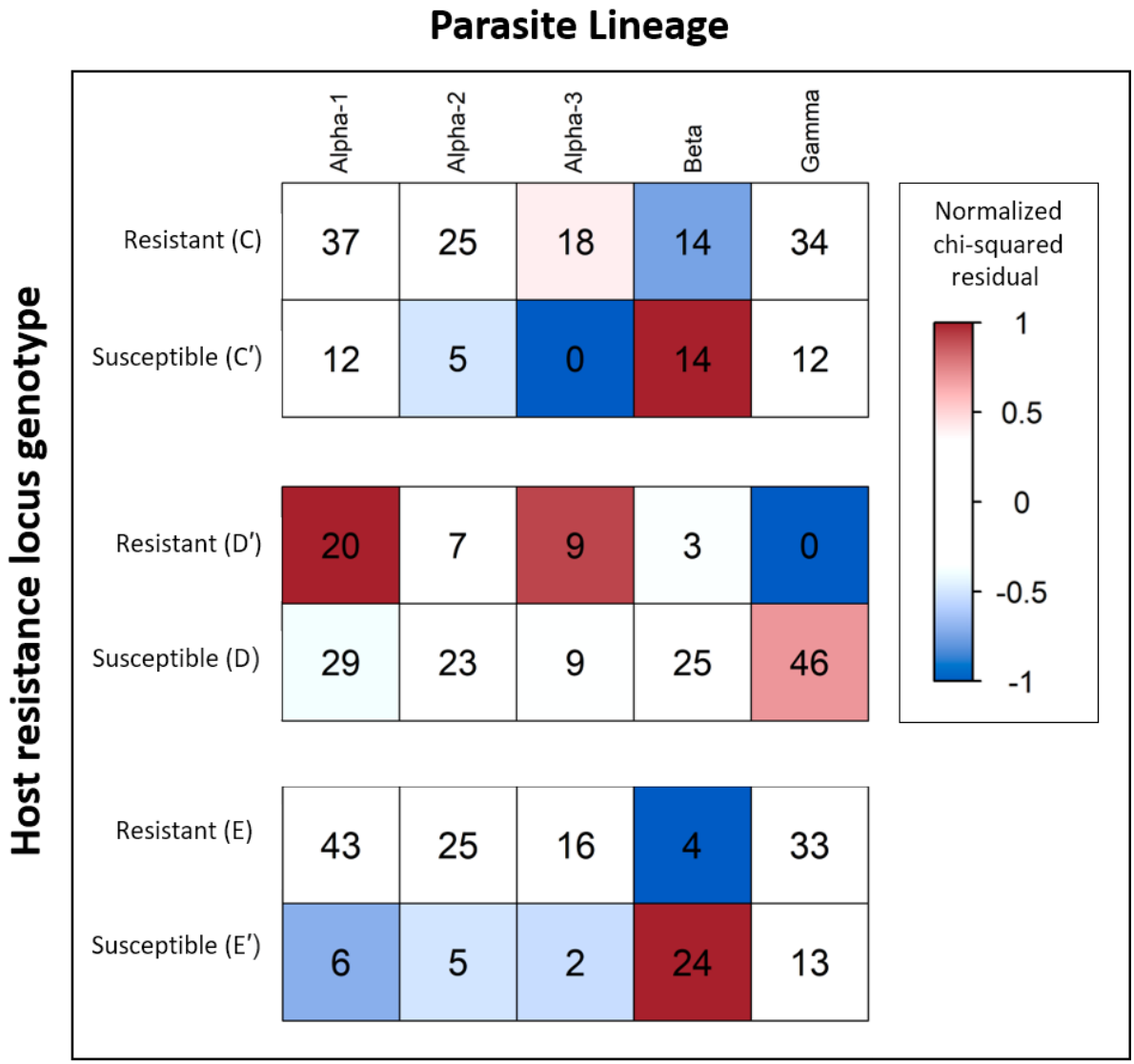


Table 1. Summary of top model results from each co-GWAS model. Rather than set a specific p-value threshold for statistical significance, the peaks described below represent the top 500 associations for each model as plotted in figure 2.

Model	Peak position (*D. magna*)	Peak description (*D. magna*)	Peak position (*P. ramosa*)	Peak description (*P. ramosa*)
1	Chromosome 4 (right arm) Contig 11: 2084705 – 2382802	Overlaps the C locus	1565875 – 1600196	Overlaps PCL triplets [9-10-11], [12-13-14], and]36-37-38]
	Chromosome 5 (right arm) Contig 71: 160694	Single SNP, intragenic, possible mapping error	758063	Aminofutalosine synthase (electron transport)
2	Chromosome 4 (right arm) Contig 11: 2084705 – 2357097	Overlaps the C locus	1565739 – 1628258	Overlaps PCL triplets [9-10-11], [12-13-14], and [36-37-38]
	Chromosome 7 (right arm) Contig 18: 1785587 – 1803392	Overlaps the D locus	318169 – 343599	Overlaps PCL triplet [21-27-28]
	Contig 32: 849103	Single SNP, intragenic, possible mapping error	1347536 – 1382775	IMP dehydrogenase and DNA topoisomerase subunit B
3	Chromosome 5 (left arm) Contig 1: 6277551 – 6506684 Contig 67: 230263 – 233072	Overlaps the E locus	1588560 – 1588747	Overlaps PCL triplet [36-37-38]
			4530 – 4707	3 Kb from PCL triplet [6-7-8]
	Chromosome 5 (right arm) Contig 25: 43625 – 1636178 Contig 53: 182869 – 443318	Shows strong correlation with resistance to *P. ramosa* isolates P20	4530 – 8506	Overlaps PCL triplet [6-7-8]
			319494 – 319584	Overlaps PCL triplet [21-27-28]
			1519089 – 1520569	Overlaps PCL triplet [24-25-26]
			1588269 – 1589748	Overlaps PCL triplet [36-37-38]
	Chromosome 7 (left arm) Contig 3: 4270730 – 4354576	Mostly uncharacterized genes	1181208 – 1205486	Overlaps PCL triplet [15-16-17]
	Chromosome 7 (right arm) Contig 18: 1785587 – 1802479	Overlaps the D locus	1516194	1 Kb distant from PCL triplet [24-25-26]