Supplementary Material

This is the supplementary material for:
O'Brien, AM, C Jack, ML Friesen, and ME Frederickson. Whose trait is it anyways?: Coevolution of joint phenotypes and genetic architecture in mutualisms. *Proceedings of the Royal Society B: Biological Sciences* **DOI: 10.1098/rspb.2020.2483**

Broader investigation of evolutionary parameters

In addition to the scenarios discussed in the main text, we also investigated more subtle contributions of particular parameters. We focused on parameters in two groups, those that affect the capacity to respond to selection (which we examine for main effects), and those that put different selection pressures on hosts and microbes (which we examine for main and interactive effects in two different sets of simulations). We ran these simulations for longer (1,000 generations) to allow simulations to reach equilibrium dynamics, and repeated each simulation 5 times with the same parameters, reporting means and variances of response measures across the 5 replicates.

Parameters that primarily affect the efficacy of selection relative to drift or the amount of genetic variance on which selection can act include distribution of fitness effects of mutational inputs, locus number in genome, population size, and gene transfer dynamics. For population size, we kept the number of hosts and microbes the same, and varied population size from 100-4100 individuals. We varied the number of loci (L) from 5-50, including changing L_M and L_H separately or together, and when manipulating together, we included twice as many microbial loci as host loci. Mutational inputs were exponentially distributed, and we explored effects of the rate parameter (λ) from 35 (smaller effect sizes of mutations) to 17 (larger effect sizes) and mutation rates (P_μ, 2×10^{-05} - 1×10^{-03}). We also varied the probability of horizontal gene transfer, P_{hrz}, from 0 to 1.

For the independent effects of parameters that affect the nature of selection, we take the same strategy as for the parameters that affect the efficacy of selection, manipulating each one at a time, and host or microbe versions of the same parameter independently. For interactive effects, we held either the host or microbe parameter constant, and factorially manipulated the remaining parameters. We manipulated the variance in the direct fitness function (ω^2), which is inversely proportional to the strength of selection via the component of fitness it is found in (i.e. determines how steeply fitness declines away from the optimum; main effects: 0.1-5, interactive: ω_H= 0.75 , ω_M= 0.25 - 1.75). Second, we changed the distribution of fitness across the direct (α) and indirect partner fitness feedbacks components ($1-\alpha$), i.e. the amount that fitness in the focal partner is determined by the distance of it's own trait value from it's own optimum, versus it's partner's trait value from the partner's optimum (main effects: 0-0.9, interactive: $\alpha_H = 0.6$, α_M = 0-1). Third, we altered the trait optima themselves (θ), which affect whether conflicts exist and the magnitude of trait space over which conflicts play out (main effects: $\theta_H = 3$ always, $\theta_M = 2$ or 3, interactive: $\theta_M = 2$ always, θ_H = 2-5). For simulations in which a parameter was not manipulated (the "background" case) parameters took the following values: $N = 2000$, $L_M = 40$, $L_H = 20$, $\lambda = 25$, $P_\mu = 10^{-4}$, $P_{hrz} = 0.2$, $\omega_H = \omega_M = 0.75$, $\alpha_H = \alpha_M = 0.6$, $\theta_M = 2$, and $\theta_H = 3$ (single manipulation only).

We summarized scenarios by a few key metrics, which we sampled at the final timestep (1,000 generations). We measured the difference (absolute value) between each species' trait optima and the average population breeding value when randomly interacting. We "measured" the standing genetic variation in each genome as the variance of the individual breeding value effects on traits of the population of either hosts or microbes. We defined fitness conflict as the realized correlation in fitness given the interacting set of hosts and microbes. We also considered terminal states in simulations for interactive parameter effects. We defined this as the rate of trait change (slope) in the last few (10) generations, either at 0 (true equilibrium) or a terminal value. Simulations are random, so will always continue fluctuating, even without conflicts.

Mechanism of parameters influencing evolutionary scenarios

Effects of direct and indirect links to fitness on evolutionary trajectories are broadly similar across hosts and microbes, and variation in the strength of partner fitness feedbacks has a large impact on outcomes. Intuitively, without partner fitness feedbacks, there is no fitness benefit of associating with a fit partner and fitness conflicts are highest. In our model, increasing fitness feedbacks pushes effective trait optima together, regardless of the θ values, inducing more positive fitness correlations and reducing ongoing conflict (Figures S6, S8). Moreover, the influence of fitness feedbacks depends on variance in the fitness functions, controlled by the ω parameters. Relatively flatter selection landscapes in a focal species (larger ωs) allow relatively weaker feedbacks (smaller αs) to shift effective optima, move phenotypes towards the partner's optimum, and induce positive fitness correlations, especially when the partner's direct relationship between trait values and fitness is steep (small ω). Reciprocally, relatively steep selection landscapes in the direct fitness link (small ωs) of a focal species can also induce positive fitness correlations with a partner, as long as the partner has an indirect fitness link ($\alpha < 1$) – because this causes the partner's effective optimum to move towards the focal population's optimum. Fitness feedbacks have the most dramatic effects on conflict when direct links in one partner are weaker (but still significant), and have slight effects when both direct links are strong and similar in magnitude (see Figures S6, S8). Thus, the evolutionary trajectory of a joint phenotype in these scenarios is largely determined by the interplay between the strength of the partner fitness feedback and the shape of the direct fitness landscape.

Further simulation results of note

Some dynamics of simulations represent ranges of parameters that we expect to be unlikely, or results that are either intuitive or more subtle, and therefore warrant less attention. Intuitively, the amount of standing additive genetic variance in both hosts and microbes scales positively with population size, number of loci, mutation rate, and the average effect size of mutations. The relationship is true both with and without conflict, but is stronger with greater conflict (where θ_M and θ_H differ by more). Indeed, previous treatment of conflicts in coevolutionary scenarios has found that parameters influencing variation inputs and strength of selection change standing genetic variation (Nuismer *et al.* 2007, Gandon & Michalakis 2002, Yamamichi *et al.* 2019).

Further, increasing the strength of selection generally decreases standing genetic variation (but not always). Generally across parameter space, when optima are identical, increasing strength of selection on the microbe direct trait-fitness component (decreasing ω_M) decreases genetic variance in both microbes and hosts via indirect positive fitness feedbacks. Instead, when optima differ, increasing strength of selection on the microbe direct component decreases V_M but increases V_H – as selection is then proportionally weaker on hosts and variants in hosts fix more slowly. When there is more genetic variation, this increases fitness variation and can strengthen fitness correlations and ongoing rates of phenotypic change are likely to be higher (but depending on direct and indirect fitness components via ω and α, Figure S6).

For example, when α moves too low, realized fitness optima of hosts and microbes "pass each other" such that hosts are more fit when trait values are at the microbe optima and vice versa (Figure S6). Similarly, intermediate values of horizontal gene transfer (P_{hrz}) slightly increases the additive genetic variance in microbial genomes relative to the highest values (the most fit alleles fix quickly) and lowest values (pure clonal interference), yet realistic scenarios for horizontal gene transfer are generally on the lower end of this scale. Finally, very high values of N, L, λ and P_μ induce high genetic variance (L, λ and P_μ also reduce the efficacy of selection on any one allele), and also positive fitness correlation – however, when variance in the trait is high, extreme values of traits will cause extreme low fitness in both hosts and microbes and create positive fitness correlations.

Interestingly, realized fitness correlations are also less predictable across replicate simulations with more extreme feedbacks, but more predictable with both the strongest and weakest selection on the direct trait-fitness component. Variable outcomes for distance from optima seem to be, unsurprisingly, a mix of the patterns for fitness correlations and additive genetic variance (Figure S7). The degree to which individual simulations differ in genetic variance largely track with main effects on genetic variance (outcomes vary more in V_A when there are more loci, individuals, mutations and intermediate horizontal gene transfer), however with rare or stronger effect mutations (or very few individuals/loci), fitness correlations also are less predictable (Figure S7).

Details of *in silico* GWAS

We simulated the rest of the neutral alleles of hosts and microbes in the same simulations as described in the main text. Neutral loci were unlinked to causal loci and each other and had no fitness effects, but mutated with the same parameters. A critical issue with using GWAS to identify loci in the genomes of multiple species is whether it requires impractically large phenotypic datasets. Phenotyping several hundred (or more) host genotypes is challenging; one approach to a multi-GWAS would require phenotyping many host genotypes paired factorially with many microbial genotypes, ideally all grown in a common environment to reduce environmental trait covariances. However, it is unlikely that real-world experiments would attempt a full-factorial design, so we instead used an experimental design split into two parts: we selected a single host genotype at random to interact with 800 microbial genotypes, and similarly selected a single microbial genotype at random to interact with 800 host genotypes. This maximizes the number of unique genotypes included in the experiment while balancing the number of host and microbe genotypes. Replicating 4 times, each half of

the experiment is 3,200 host-microbe units (or 6,400 experimental units in total), a size on par with GWAS of host-pathogen interactions (e.g. Malaria Genomic Epidemiology Network, 2014), but eclipsed by sizes of GWAS for other phenotypes (e.g. Buckler *et al.* 2009). However, this design is larger than many GWAS of host-microbe mutualisms (Stanton-Geddes *et al.* 2013). Phenotypes are the additive value of the combined effects of host and microbial loci at all alleles, plus experimental error that is normally distributed with mean 0 and standard deviation $\theta_{exp} = 0.005$.

Each host and microbe genotype used in the in silico GWAS experiment was extracted after 300 simulation generations. Simulations used an infinite alleles model, yet GWAS requires biallelic SNPs, so we model our quantitative alleles as discrete, yet fully linked, loci within a region equal to the number of unique alleles, i.e., each allele is a distinct haplotype. Retaining this linkage as a source of error in estimating loci effects is desirable because linked causal variants are a major driver of error in real GWAS (Rockman, 2012). We next converted our simulated genotype and phenotype data into files formatted for GWAS software GEMMA (Zhou & Stephens, 2014) with PLINK (version 1.07, Purcell *et al.* 2007). Finally, we ran a basic linear model in GEMMA, and selected SNPs with the 5% most extreme p-values.

References

Buckler ES, Holland JB, Bradbury PJ, Acharya CB, Brown PJ, Browne C, Ersoz E, Flint-Garcia S, Garcia A, Glaubitz JC et al.. 2009 The genetic architecture of maize flowering time. *Science*, 714–718.

Gandon S, Michalakis Y. 2002 Local adaptation, evolutionary potential and host–parasite coevolution: interactions between migration, mutation, population size and generation time. *Journal of Evolutionary Biology* **15**, 451–462.

Malaria Genomic Epidemiology Network, Rockett KA, Clarke GM, Fitzpatrick K, Hubbart C, Jeffreys AE, Rowlands K, Craik R, Jallow M, Conway DJ et al.. 2014 Reappraisal of known malaria resistance loci in a large multicenter study. *Nat Genet* **46**, 1197–1204.

Nuismer SL, Ridenhour BJ, Oswald BP. 2007 Antagonistic coevolution mediated by phenotypic differences between quantitative traits. Evolution **61**, 1823–1834.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, De Bakker PI, Daly MJ et al.. 2007 PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American journal of human genetics* **81**, 559–575.

Rockman MV. 2012 The QTN program and the alleles that matter for evolution: all that's gold does not glitter. *Evolution* **66**, 1–17.

Stanton-Geddes J, Paape T, Epstein B, Briskine R, Yoder J, Mudge J, Bharti AK, Farmer AD, Zhou P, Denny R et al.. 2013 Candidate genes and genetic architecture of symbiotic and agronomic traits revealed by whole-genome, sequence-based association genetics in Medicago truncatula. *PloS one* **8**, e65688.

Yamamichi M, Lyberger K, Patel S. 2019 Antagonistic coevolution between multiple quantitative traits: Matching dynamics can arise from difference interactions. *Population Ecology* **61**, 362–370.

Zhou X, Stephens M. 2014 Efficient algorithms for multivariate linear mixed models in genome-wide association studies. *Nature methods* **11**, 407.

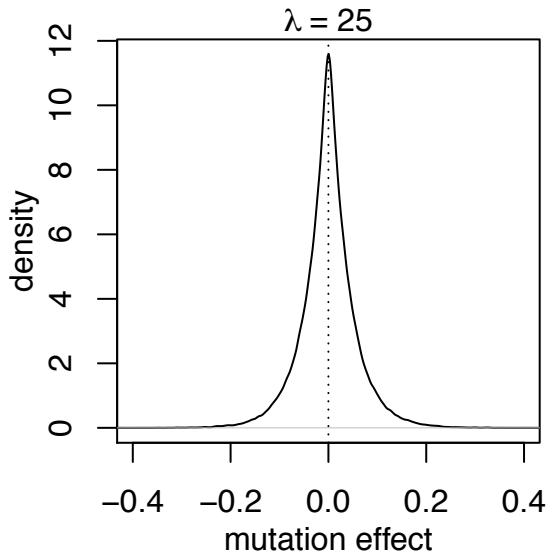

Figure S1: Example distribution from which mutation effect sizes are drawn. The depiction is cut at -0.4 and 0.4, but larger effect sizes may be drawn with increasingly lower probability. $\lambda = 25$ is used for all examples in the main text, effects of varying λ can be seen in Figure S6.

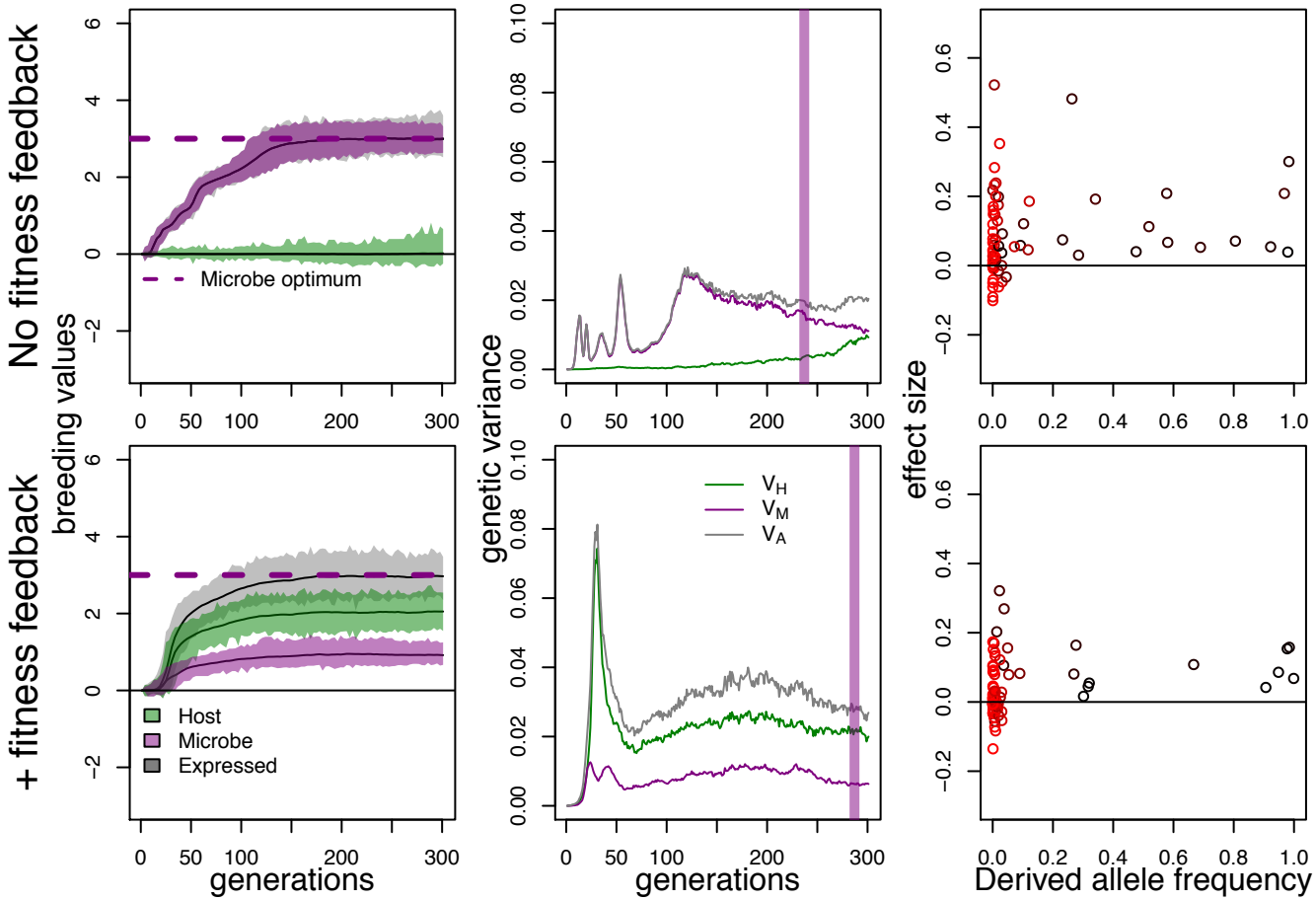

Figure S2: Results for reversing host and microbe parameters for scenarios 1 and 4, with the columns now showing the equivalents of Figures 2-S11 (see respective captions for description). Only microbes now have direct fitness consequences for the trait value. With fitness feedbacks, hosts evolve faster and retain more genetic variance, which is the same as in scenario 4 (see main text for discussion of host and microbe genetic variance in the main text). Effect size and age of segregating loci in microbes now resemble patterns for hosts from scenarios 1 and 4, Figure S12, note different scale.

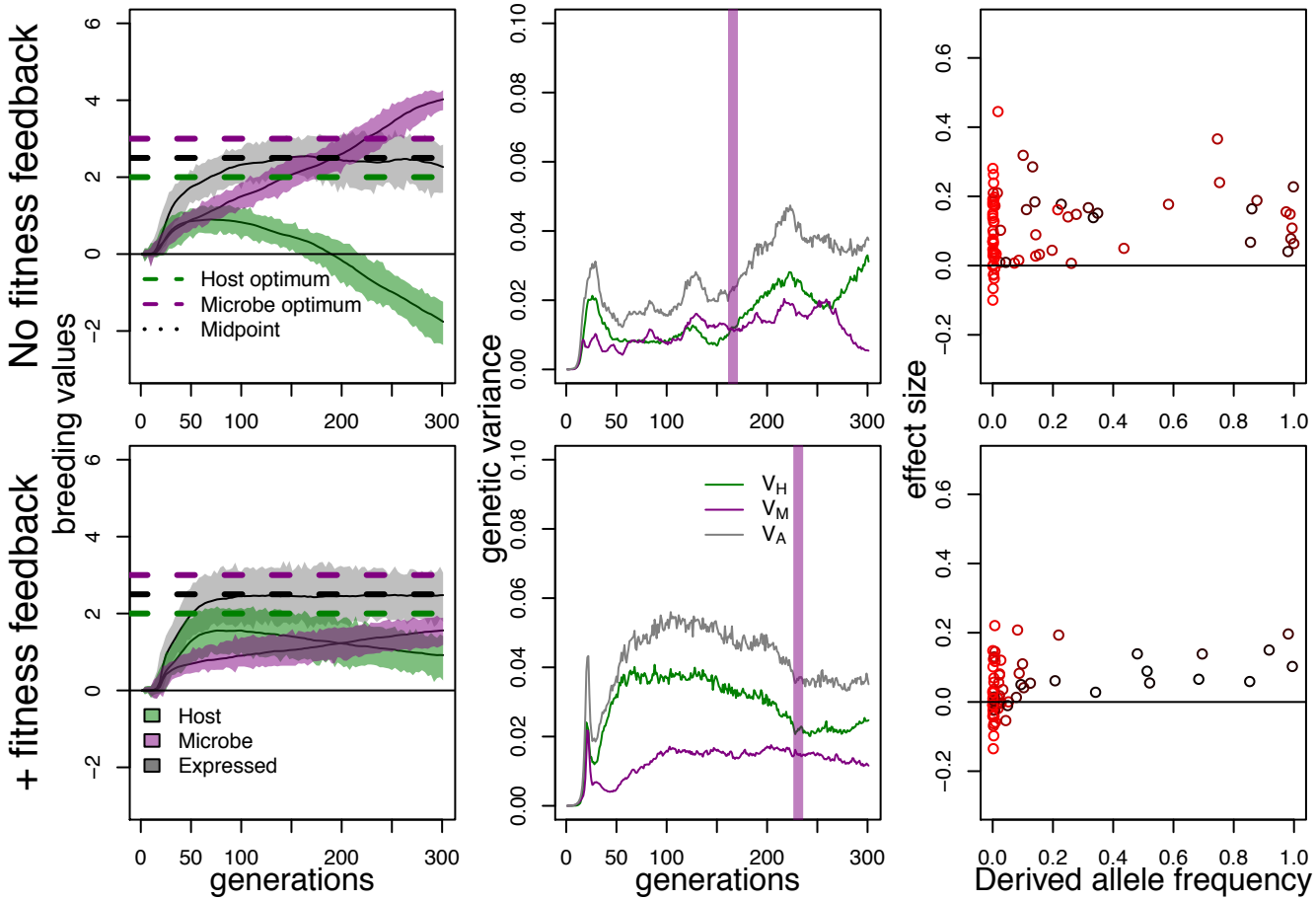

Figure S3: Results for reversing host and microbe parameters for scenarios 3 and 6, with the columns now showing the equivalents of Figures 2-S11 (see respective captions for description). The microbe optimum now exceeds the host optimum. Hosts still evolve faster and retain more genetic variance, which is the same as in scenario 4 (see discussion of genetic variance of hosts and microbes in the main text). Effect size and age of segregating loci in microbes now resemble patterns for hosts from scenario 3 and 6, Figure S12, note different scale.

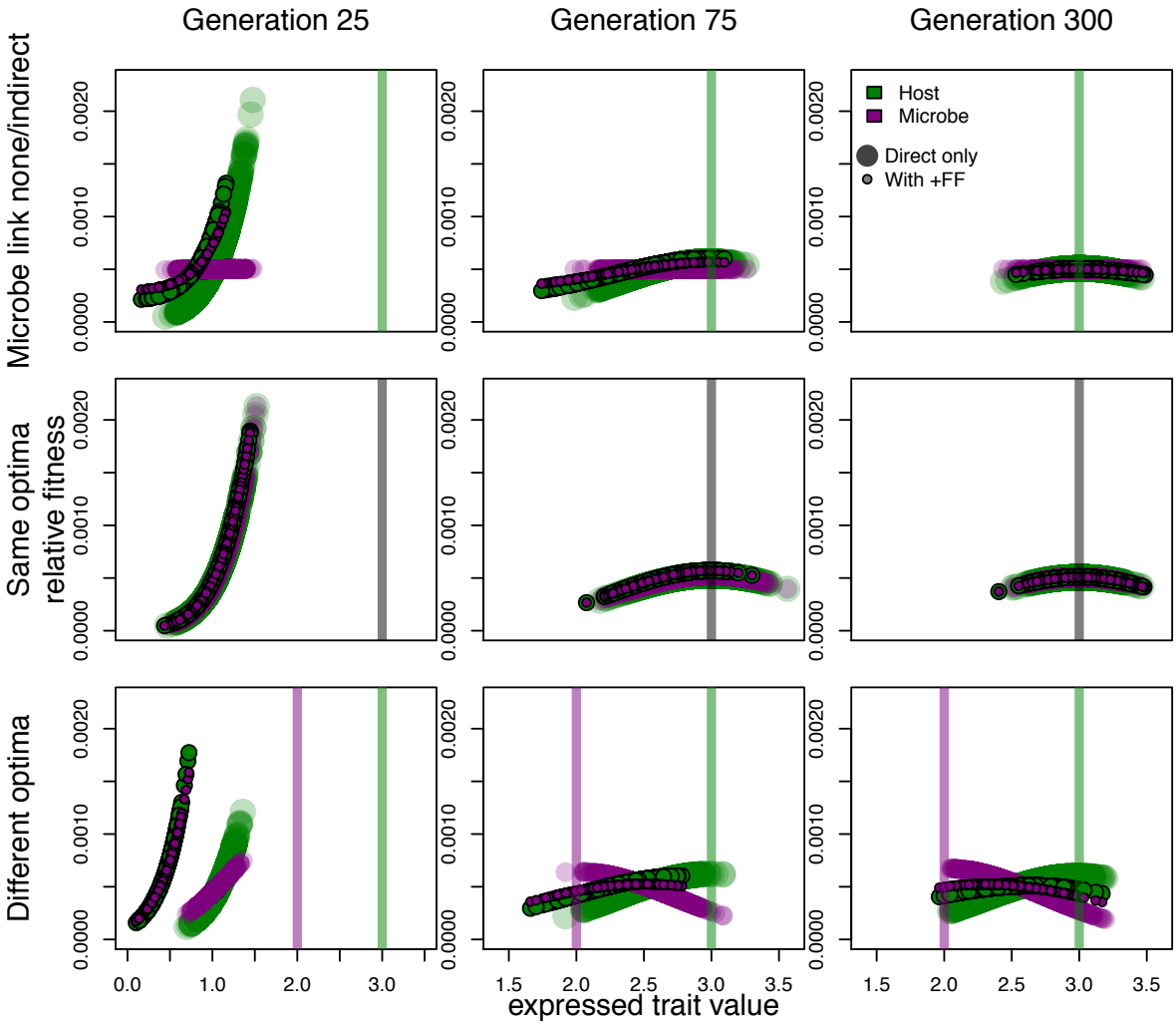

Figure S4: Trait-fitness functions over time (plots left to right) in the six scenarios, arranged top to bottom. Relative fitness on the y-axis, trait values of individuals in the simulations on the x-axis. For each trait value, both host and microbe fitness are shown. Top row: trait directly linked to host fitness only. Middle row: trait directly linked to host and microbe fitness, and they have the same optimal trait value. Bottom row: trait directly linked to both microbe and host fitness, at distinct optima. In large points with no outline, all links to traits are direct only (scenarios 1-3). In smaller points with outlines, microbe and host fitnesses are also linked via positive fitness feedbacks (scenarios 4-6). Microbe (purple), host (green), or both (grey) optima are marked with vertical lines.

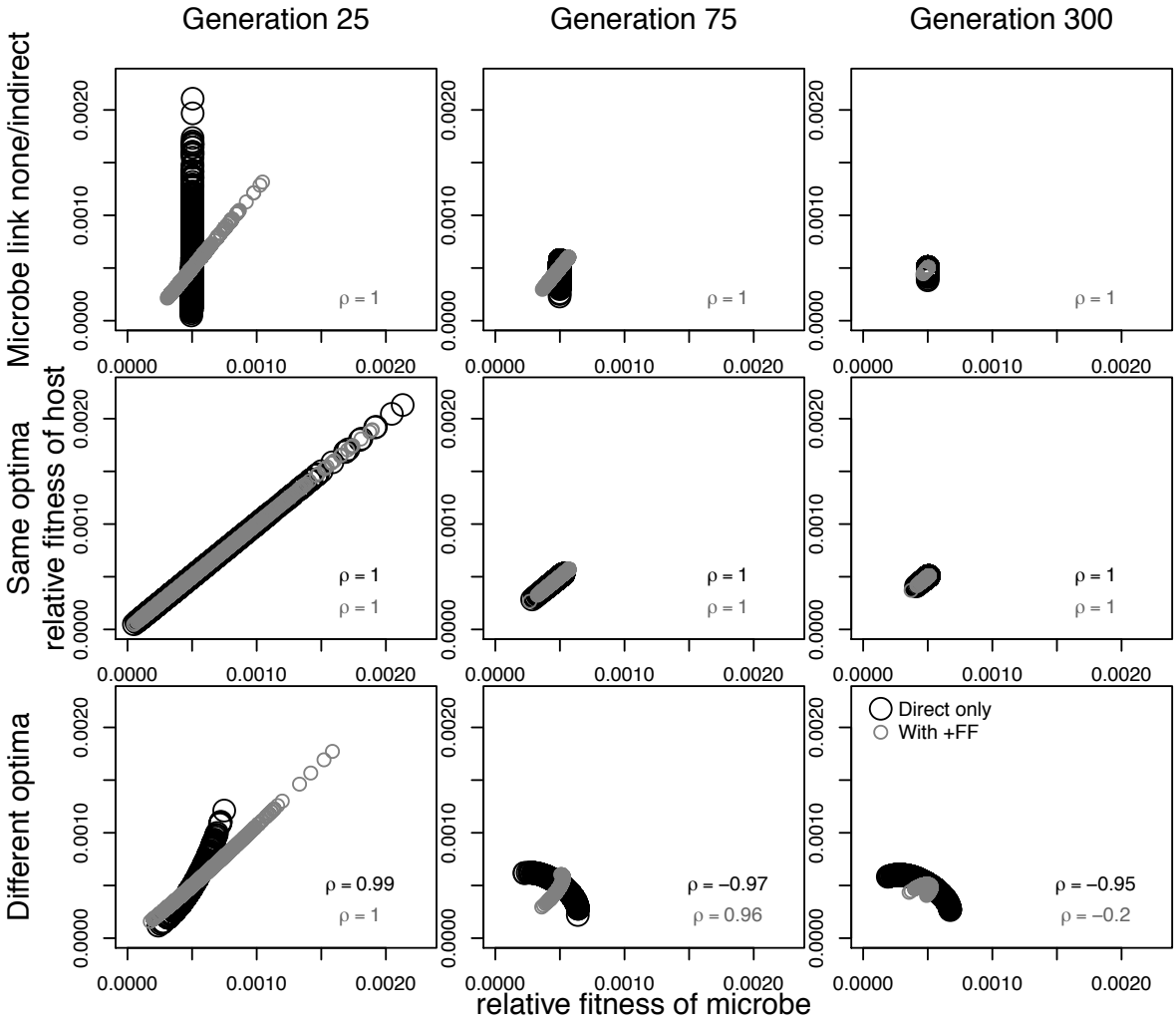

Figure S5: Fitness correlations over time (plots left to right) in the six scenarios, arranged top to bottom. Panels the same as in Figure S4, but with host fitness on the y-axis and microbe fitness on the x-axis. In black points, all links to traits are direct only (scenarios 1-3). In gray points, microbe and host fitnesses are linked via positive fitness feedbacks (scenarios 4-6). When the direct link to fitness is strong only for hosts, or when hosts and microbes have identical optima, fitness remains perfectly positively correlated or not relevant through time (top and middle rows). When trait optima differ, and are linked to the trait for both partners, initially fitness correlations are strongly positive, as both feedbacks and optima select for trait shifts in the same directions, but they can become negative or non-linear over time.

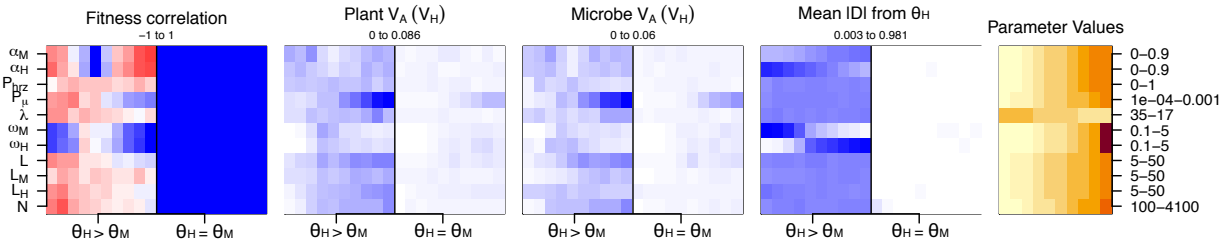

Figure S6: Parameter sensitivity survey for all parameters, average results. Each response measurement is a separate plot for the first four plots, and the value of each cell is the final timestep of the simulations averaged across the 5 replicate simulations. Left halves of plots show results when trait optima differ, and right sides when trait optima are identical. In the rightmost plot, we illustrate the values of parameters used, this grid corresponds to each half of the other four plots - note that not all parameters change linearly or positively (e.g. λ decreases, representing increasing mutation effect sizes). Only one parameter changes at a time, the rest are held constant, see "Broader investigation of evolutionary parameters". Redder color indicates more negative values. Blue indicates values are greater than 0, and bluer color indicates more positive values. Ranges below plot titles indicate the numeric scale from red to blue (fitness correlation), or white to blue (others). For L, number of host loci (legend value) is manipulated simultaneously with microbe number of loci, which is double. Loci number is manipulated separately in L_M and L_H rows. V_A refers to additive genetic variance. $|D|$ from θ_H refers to the absolute value of the distance of the expressed trait value from the host optima.

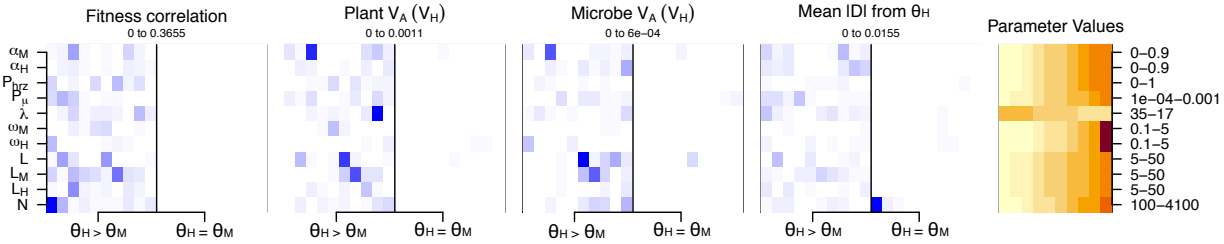

Figure S7: Parameter sensitivity survey for all parameters, variability in results. Each response measurement is a separate plot, and the value of each cell is the variance of the final timestep of the simulations across the 5 replicate simulations. Bluer color indicates more positive values, and ranges below plot titles indicate the numeric scale from white to blue. Note that the scale is very small relative to mean values for all except fitness correlation (Figure S6). Left halves of plots show results when trait optima differ, and right sides when trait optima are identical. In the rightmost and final plot, we show the unique values of parameters used, this grid corresponds to each half of the other four plots. Only one parameter changes at a time, the rest are held constant. Note that λ decreases, but this represents increasing mutation effect sizes. For L, number of host loci (legend value) is manipulated simultaneously with microbe number of loci, which is double. Loci number is manipulated separately in L_M and L_H rows.

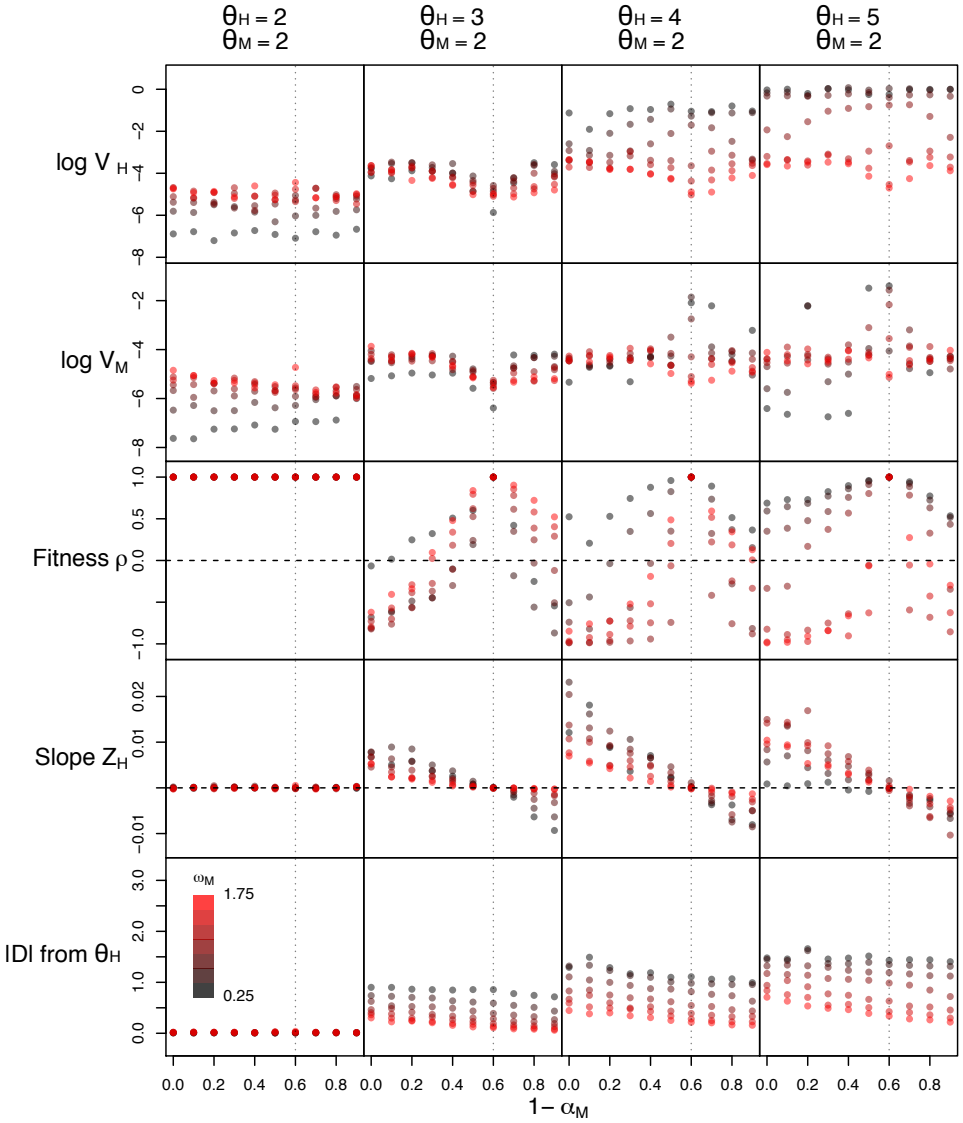

Figure S8: Parameter sensitivity survey for combinations of trait optima, direct trait-fitness links and contributions of partner fitness feedbacks. Each set of parameters is represented by one point, which is an average of the response variable at the final timestep (1,000, or final 10 timesteps for slopes) across 5 replicate simulations, plotted against $1 - \alpha_M$ (increasing partner fitness feedback). $\alpha_H = 0.6$ in all simulations. Each row of plots shows a different response variable. Redder colors indicate greater ω_M ($\omega_H = 1$ in all). The dashed vertical line in each plot represents an important switch to extreme feedbacks: to the left of this line, the partner fitness feedback exceeds the direct contribution to fitness and the microbe inclusive fitness is maximal closer to θ_H than where hosts fitness is maximized. In other words, the dashed vertical line is where the trait-fitness peaks of hosts and microbes 'pass each other.' ρ is the fitness correlation. Slope Z_H is the slope of the host breeding value. Each column of plots shows a different combination of trait optima (from highest to lowest distance between optima, $\theta_M = 2$ in all).

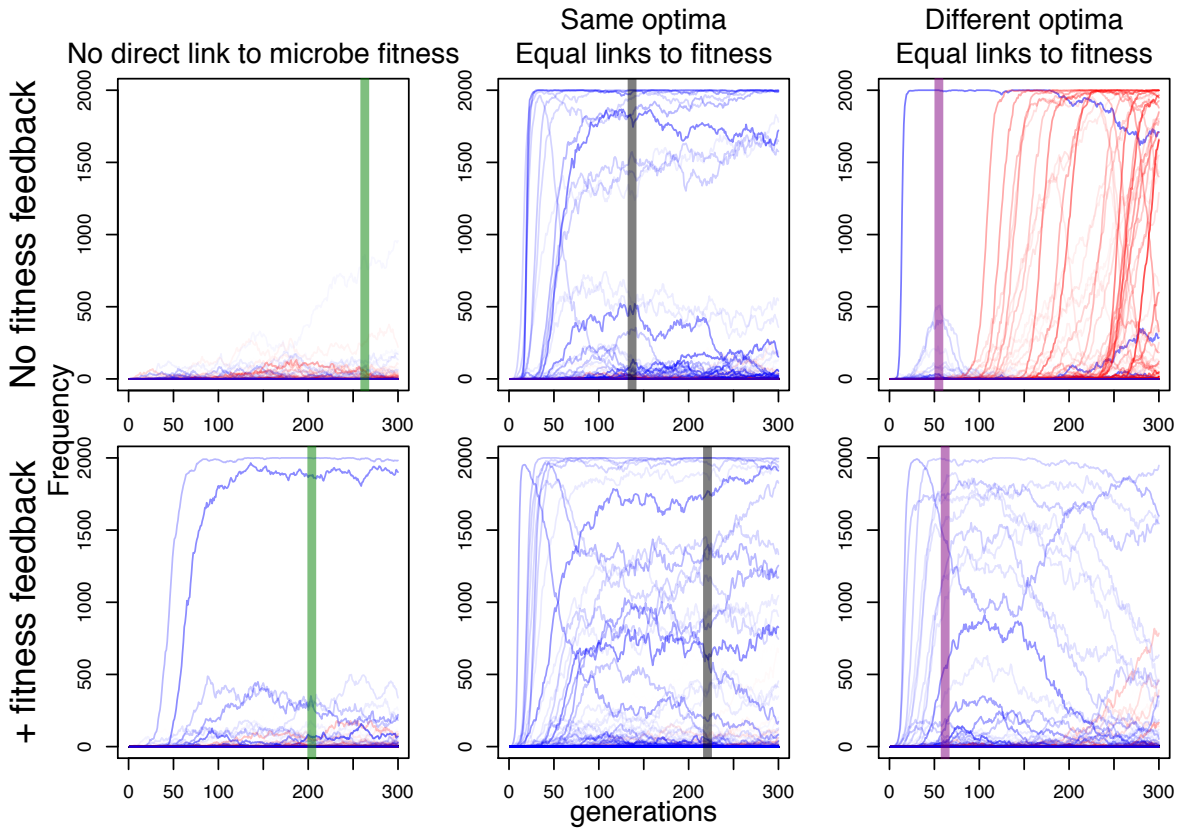

Figure S9: Allele trajectories through time for all microbial loci for all scenarios. Alleles with positive effects on the phenotype are in blue, negative effects in red. Alleles with larger effect sizes are darker colors. As in Figure 3, vertical green (host), gray (both), purple (microbe) lines on each panel mark when the average trait value reaches (or passes, panels on the right) the respective trait optimum. Top row: left, trait directly linked to host fitness only. Middle, trait directly linked to host and microbe fitness; host optimum = microbe optimum. Right, trait directly linked to both microbe and host fitness, at separate optima. Bottom row: as above, but host and microbe fitnesses are linked via positive fitness feedbacks.

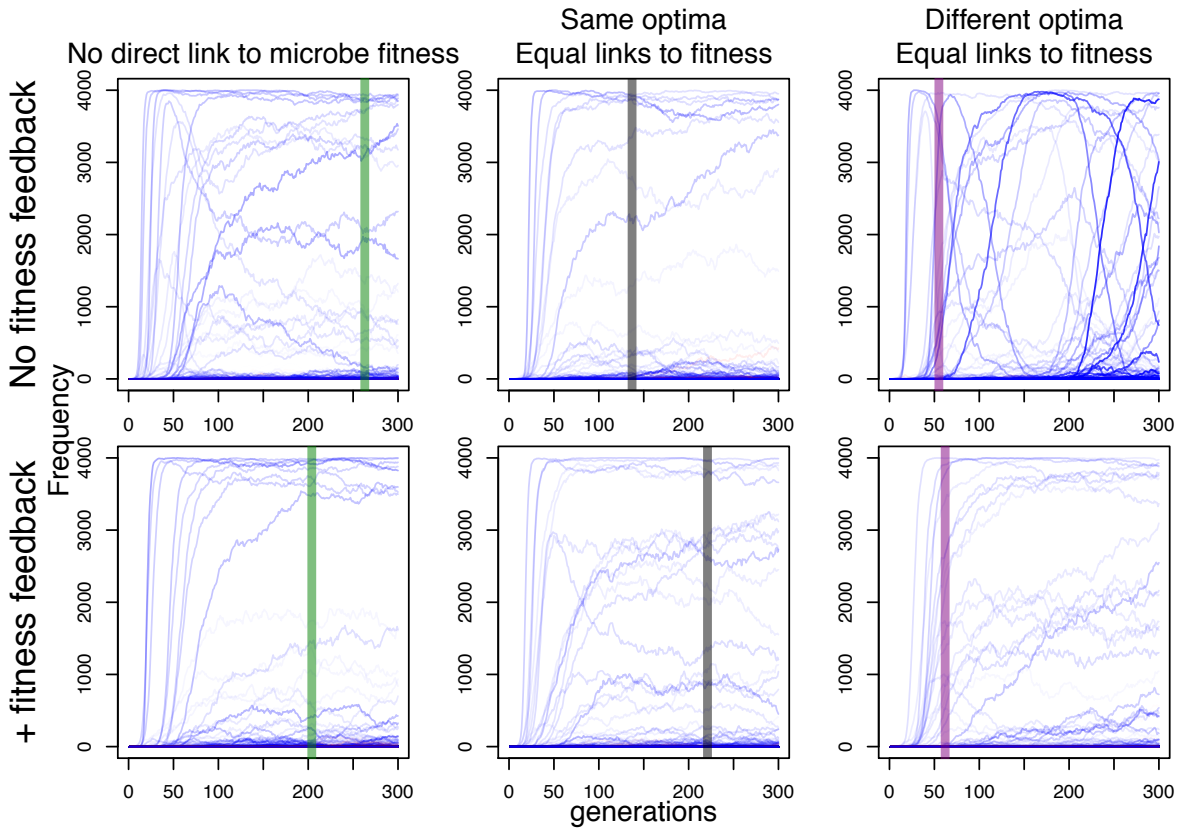

Figure S10: Allele trajectories through time for all host loci for all scenarios. Alleles with positive effects on the phenotype are in blue, negative effects in red. Alleles with larger effect sizes are darker colors. As in Figure 3, vertical green (host), gray (both), purple (microbe) lines on each panel mark when the average trait value reaches (or passes, panels on the right) the respective trait optimum. Top row: left, trait directly linked to host fitness only. Middle, trait directly linked to host and microbe fitness; host optimum = microbe optimum. Right, trait directly linked to both microbe and host fitness, at separate optima. Bottom row: as above, but host and microbe fitnesses are linked via indirect positive fitness feedbacks.

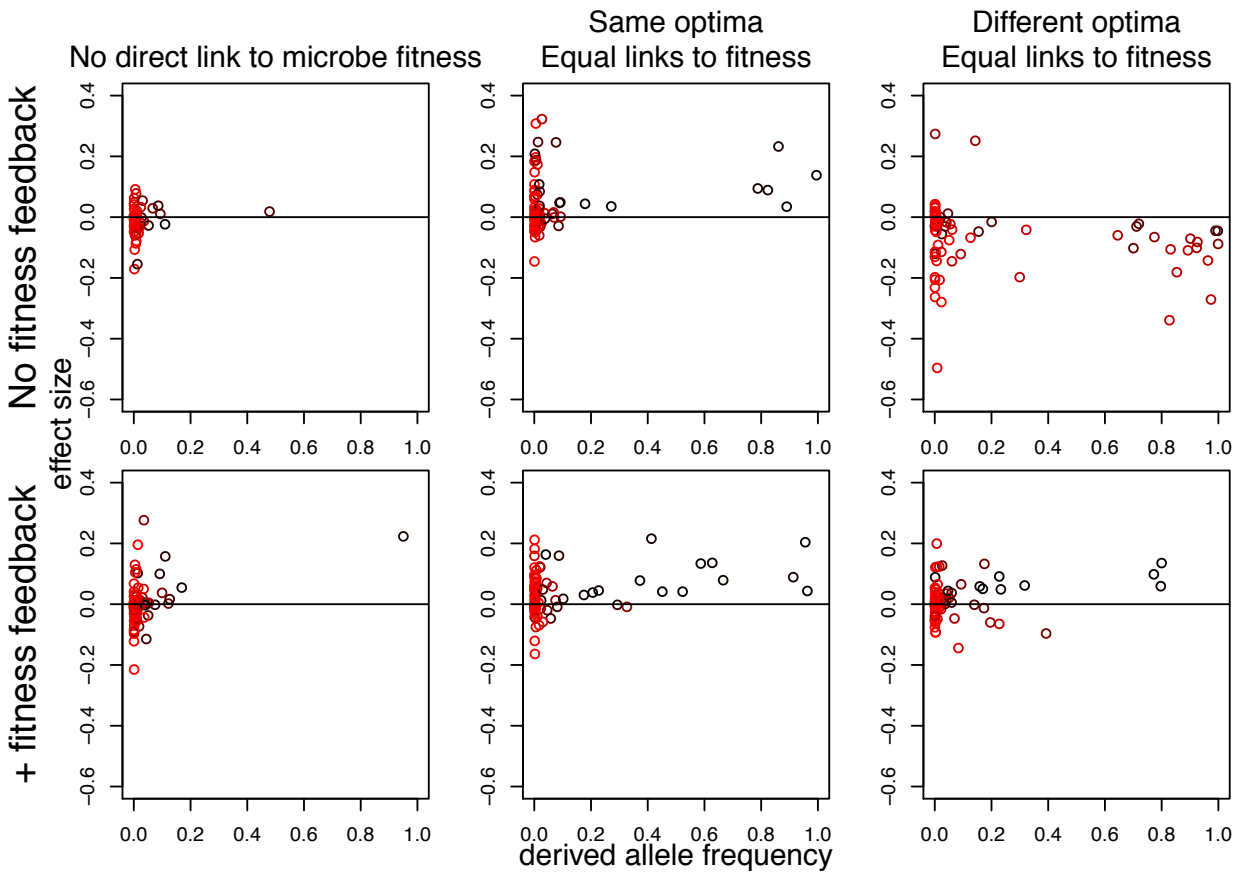

Figure S11: Effect size, sign, and allele frequency distribution for segregating derived alleles in microbes at the final generation, with color indicating allele age (redder indicates more recently derived alleles) for each scenario in Figure 2.

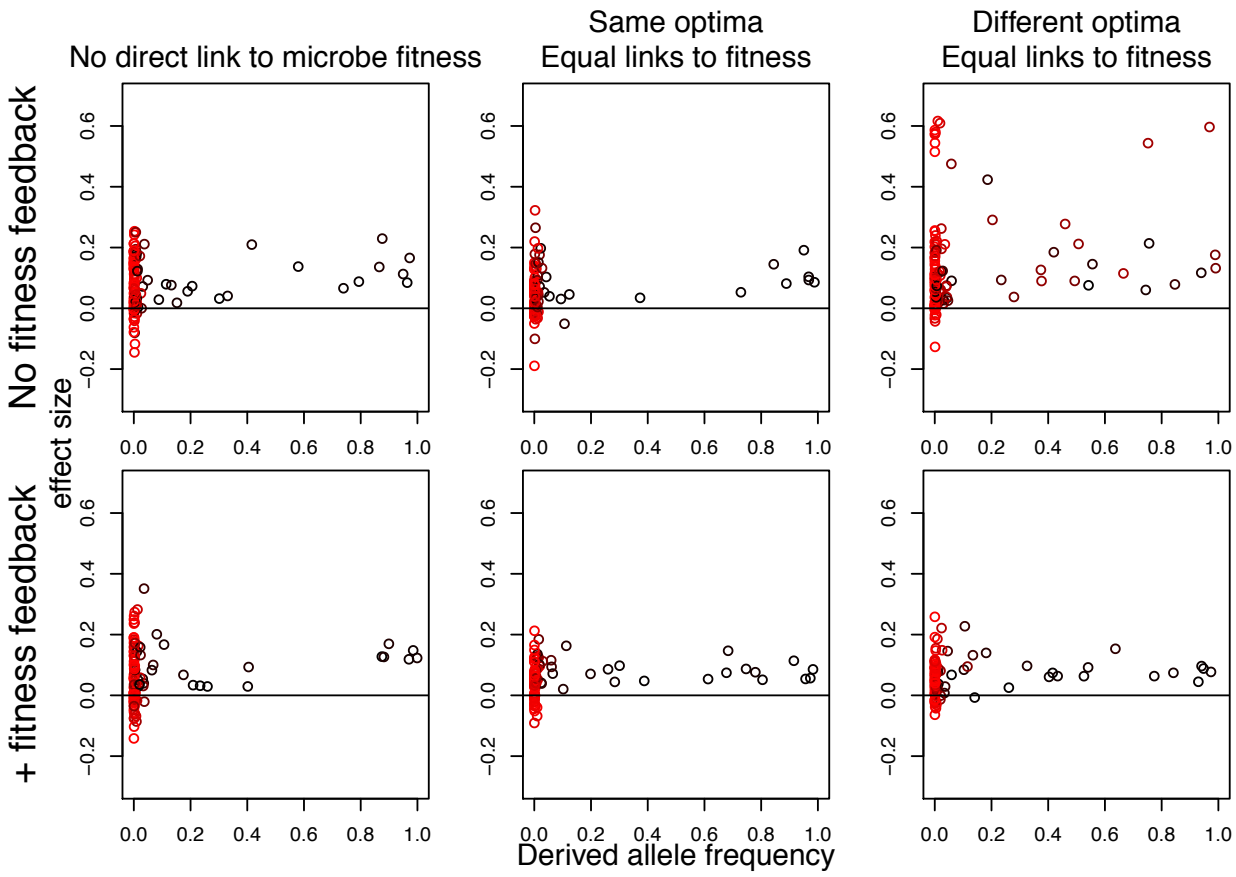

Figure S12: Effect size, sign, and allele frequency distribution for segregating derived alleles in the host at the final generation, with color indicating allele age (redder indicates more recently derived alleles) for each scenario in Figure 2.

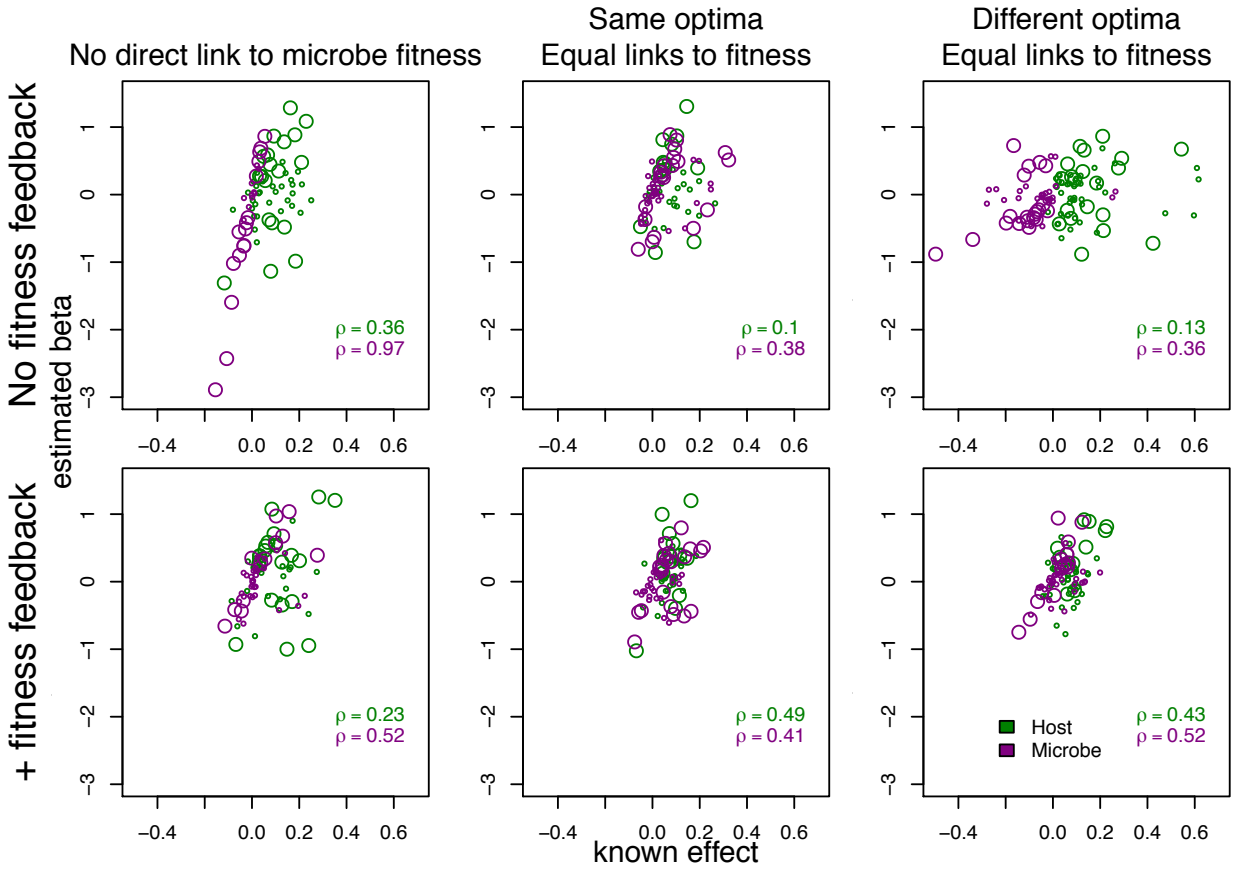

Figure S13: GWAS estimated betas (y-axis) are correlated to known simulated allele effects (x-axis), for both host (green) and microbe (purple) alleles. Loci detected as significant by GWAS are shown in larger plotting symbols than loci not detected as significant. Panels are the same scenarios as in Figure 2.

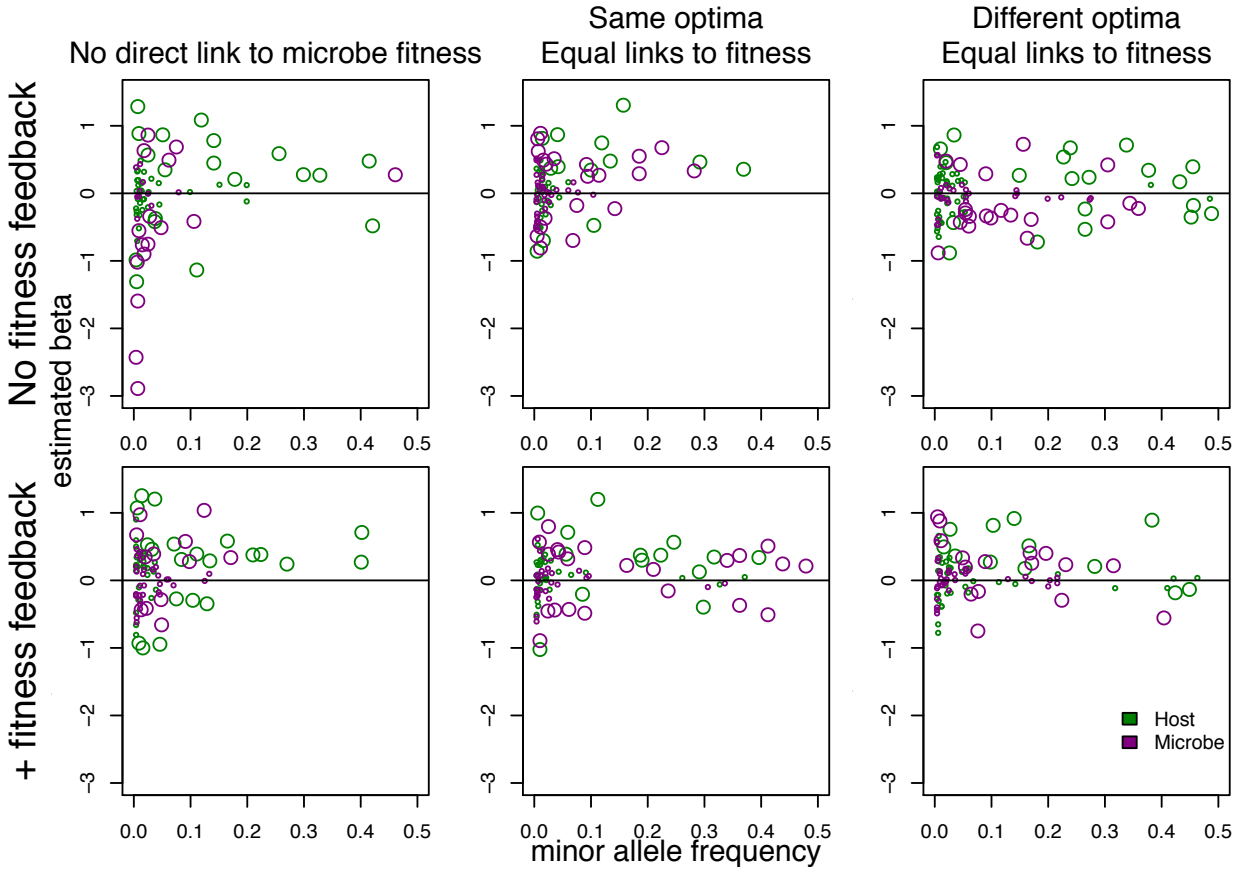

Figure S14: GWAS is better at detecting loci with larger effects (y-axis) relative to the experimental mean and higher minor allele frequencies (x-axis, among the loci in the experiment), for both host (green) and microbe (purple) loci. Loci detected as significant by GWAS are shown in larger plotting symbols than loci not detected as significant. Panels are the same scenarios as in Figure 2.

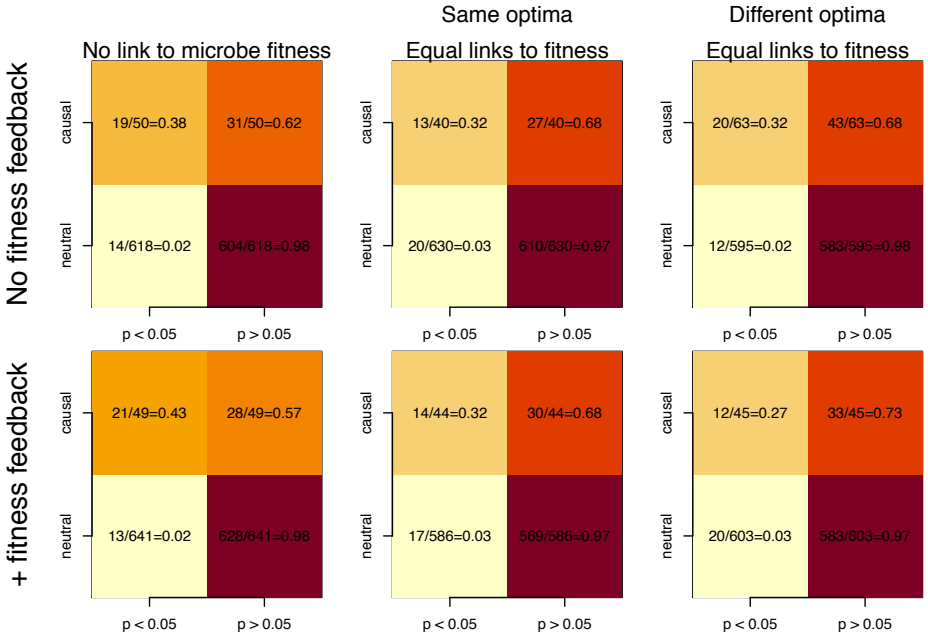

Figure S15: True and false positive rates for GWAS of simulated host loci to *in silico* phenotypes in all six scenarios, as in Figure 2. Each panel shows the number of causal (top squares) and neutral (bottom squares) loci GWAS detects as significant at $p < 0.05$ (left squares), or does not detect ($p > 0.05$, right squares). Each detected number is divided by the total number of causal (top squares) and neutral (bottom squares) loci (denominators), to produce true and false positive rates (left squares) and true and false negative rates (right squares). Higher rates are colored redder on a scale from 0 (white) to 1 (dark red). Identified loci include an appreciable amount of causal loci.

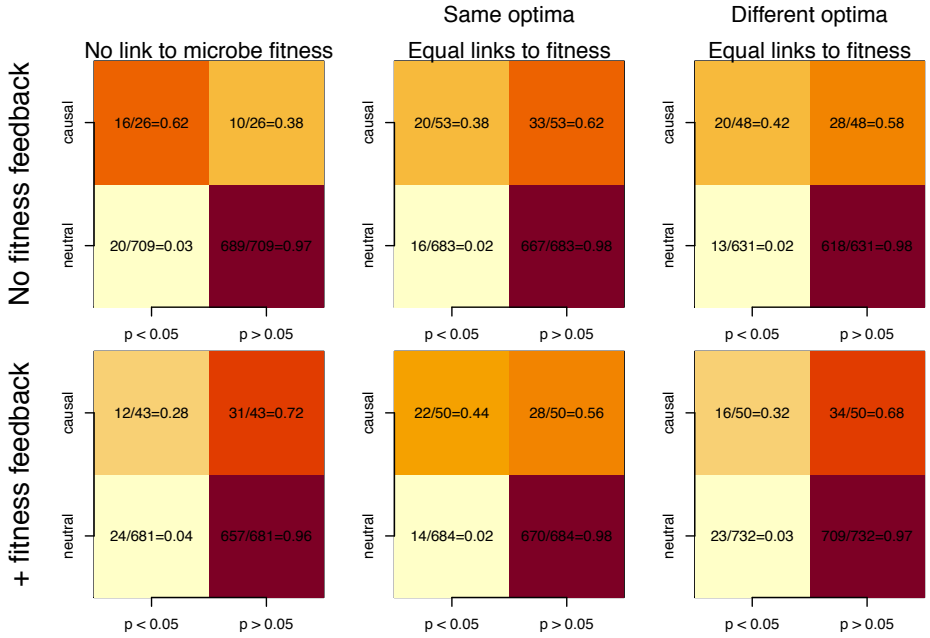

Figure S16: True and false positive rates for GWAS of simulated microbe loci to *in silico* phenotypes in all six scenarios, as in Figure 2. Each panel shows the number of causal (top squares) and neutral (bottom squares) loci GWAS detects as significant at $p < 0.05$ (left squares), or does not detect ($p > 0.05$, right squares). Each detected number is divided by the total number of causal (top squares) and neutral (bottom squares) loci (denominators), to produce true and false positive rates (left squares) and true and false negative rates (right squares). Higher rates are colored redder on a scale from 0 (white) to 1 (dark red). Identified loci include an appreciable amount of causal loci.