Bacterial lifestyle shapes pangenomes

Supplementary Material 1

Abstract

This document contains full details of all of our analyses, including those we refer to in the main text and additional supporting analyses. We have organised the documents by result, and include full details of the methods for each section. This document is also available to downloads as both an R markdown file and an R script, to enable full access to all our code. Data and the phylogeny are also provided. It is split into the following parts:(1)Pangenome fluidity correlates with whether a species is host-associated or free living;(2)Within hosts, which of four key lifestyle traits matter for shaping pangenome fluidity?(3)What about other factors in addition to lifestyle: genome size and effective population size?(4)Other measures of pangenome variation across species.

Contents

1 Pangenome fluidity correlates with whether a species is host-associated or free living.

One major lifestyle trait is whether bacteria live freely or inside hosts. It is likely that species which live freely encounter more opportunities for gene gain, compared to host-associated species. Additionally, free-living species will likely encounter more environmental variability than host-associated species, meaning each environment may exert different selection pressures, requiring different sets of genes of the individuals present there. Together, these factors might cause pangenome fluidity to be higher in species which live freely, and lower in species which are associated with hosts.

To test this prediction, we categorised species into host-associated or free-living in several ways.

1.1 Host-associated, Free-Living, or Both

First, we categorised species into whether they were 'Host-associated', 'Free-living' or 'Both'. We then assessed whether this was correlated with pangenome fluidity. We did this across 125 species which we could categorise into one of 'host', 'both' or 'free'.

1.1.1 Linear model

Using a linear model, we found that host-association was significantly correlated with species' pangenome fluidity: 'Host' species had lower pangenome fluidity than both species which were 'free-living' and species which were sometimes 'host-associated' and sometimes 'free-living' (values for rows 2 and 3 are testing if there is a difference in the pangenome fluidity of species which were 'Both' and 'Free', compared to 'Host').

Table S1: Results from a linear model with pangenome fluidity as the response variable and host-associated, free-living or both as the explanatory variable.

	Estimate	Std.Error	t.value	p.value	signif.
(Intercept)	0.15546	0.01032	15.062	0.0000	***
Host_or_freeBoth	0.08204	0.01454	5.644	0.0000	***
Host_or_freeFree	0.12632	0.04129	3.060	0.0027	**

1.1.2 MCMCglmm

However, this approach treats all species as independent data points, which is not the case due to shared ancestry. Therefore, we next ran a MCMCglmm model to control for phylogenetic relationships between species. We did this within the MCMCglmm by incorporating a matrix-version of the phylogeny as a random effect.

To see how we did this, see code below:

```
## Label tree nodes with numbers
# 'dataTree' is our phylogeny as an ultrametric tree in nexus format
dataTreeNode <- makeNodeLabel(dataTree, method = "number")

## Make matrix of tree called 'Ainv'
INtree <- inverseA(dataTreeNode, nodes="TIPS") # Converts phylogeny into a covariance matrix
Ainv <- INtree$Ainv #Extracts covariance values

# Weakly informative prior
prior <- list(R=list(V = 1, nu = 0.002), G=list(G1=list(V=1,nu=0.002)))

mcmc_model_1 <- MCMCglmm(pangenome_fluidity ~ Host_or_free, random=~Species,
                         data=pangenome_lifestyles_host_free_no_unknown,
                         nitt=50000,
                         prior=prior, ginverse = list(Species = Ainv),verbose = FALSE)
```

Our MCMCglmm analyses produced similar results: host-associated species have lower pangenome fluidity than both species which are sometimes free-living ('Both') and species that are always free-living ('Free'). We found that once accounting for phylogeny, which explained approximately 48% of the variance in pangenome fluidity, whether a species was host-associated or free-living explained around 14% of the variance in pangenome fluidity.

Table S2: Results from the above MCMCglmm with pangenome fluidity as the response variable, whether a species is host-associated, free-living or both as the fixed effect, and phylogeny as a random effect.

	post.mean	l-95% CI	u-95% CI	eff.samp	pMCMC	signif.
(Intercept)	0.17858	0.10804	0.2488	4700	0.0002128	***
Host_or_freeBoth	0.07653	0.04727	0.1070	4700	0.0002128	***
Host_or_freeFree	0.11931	0.04048	0.1971	4265	0.0034043	**

	R-squared value
Fixed effect	0.1380
Random effect	0.4840
Total model	0.6221

We can view these results together with the data in Figure S1.

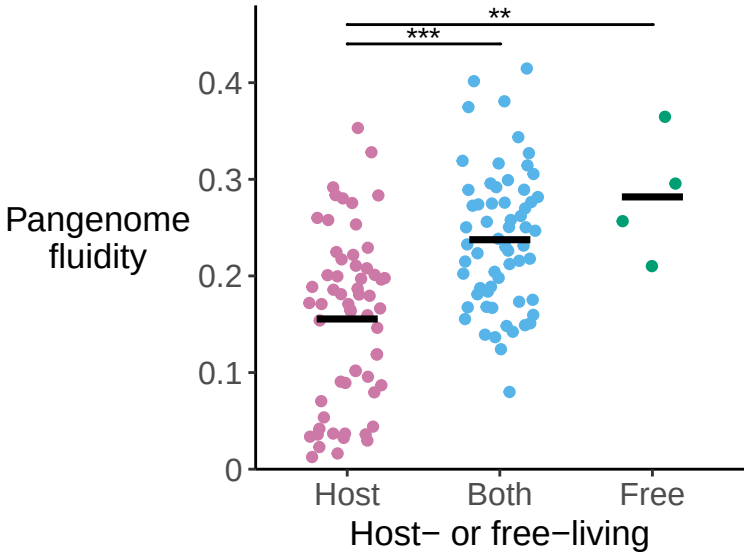


Figure S1: Host-association correlates with pangenome fluidity.

1.2 Primary environment: Host-associated or Free-living

For the previous method, species' in the 'both' category are treated as a separate category, even though some may be predominantly host-associated or free-living, and only rarely live in the other. To simplify this, we recorded each species' primary environment and categorised this as either 'Host-associated' or 'Free-living'. Please see the table below to see some examples of this:

Table S3: Example of species' primary environments, and their 'Host' or 'Free' catgeories.

Species	Primary_environment	Category_primary_env
Acetobacter_pasteurianus	Plant	Host
Acinetobacter_baumannii	Human	Host
Acinetobacter_pittii	Human	Host
Aeromonas_hydrophila	Water	Free
Alteromonas_mediterranea	Marine	Free
Bacillus_amyloliquefaciens	Soil	Free

1.2.1 Linear model

We found with a linear model that species with a free-living primary environment had significantly higher pangenome fluidity than those with a host-associated primary environment.

Table S4: Results from a linear model with pangenome fluidity as the response variable and whether a species' primary environment is host-associated or free-living as the explanatory variable.

	Estimate	Std.Error	t.value	p.value	signif.
(Intercept)	0.18824	0.009185	20.49	0.0000	***
Category_primary_envFree	0.04411	0.018153	2.43	0.0165	*

1.2.2 MCMCglmm

We also found the same when we ran an analogous MCMCglmm to control for phylogenetic relationships between species: free-living species have a higher pangenome fluidity than host-associated species, although the R-squared of this method of categorising species' (Fixed effect row in Table 3) was quite low. This suggests the 'Both' category in the previous section captures some extra variance in pangenome fluidity, even though species within that category are likely highly variable.

Note: this MCMCglmm was formatted as in the code above, just with a different fixed effect. All MCMCglmm models in the rest of this document are also formatted as in that code, unless specified. Full code for all models is available at *ADD LINK HERE LATER*.

Table S5: Results from a MCMCglmm with pangenome fluidity as the response variable, whether a species' primary environment is host-associated or free-living as the fixed effect, and phylogeny as a random effect.

	post.mean	l-95% CI	u-95% CI	eff.samp	pMCMC	signif.
(Intercept)	0.20427	0.130715	0.28843	4113	0.0002128	***
Category_primary_envFree	0.03589	0.001864	0.07076	4700	0.0434043	*

	R-squared value
Fixed effect	0.01945
Random effect	0.56025
Total model	0.57970

We can view these results with the accompanying data in the figure below.

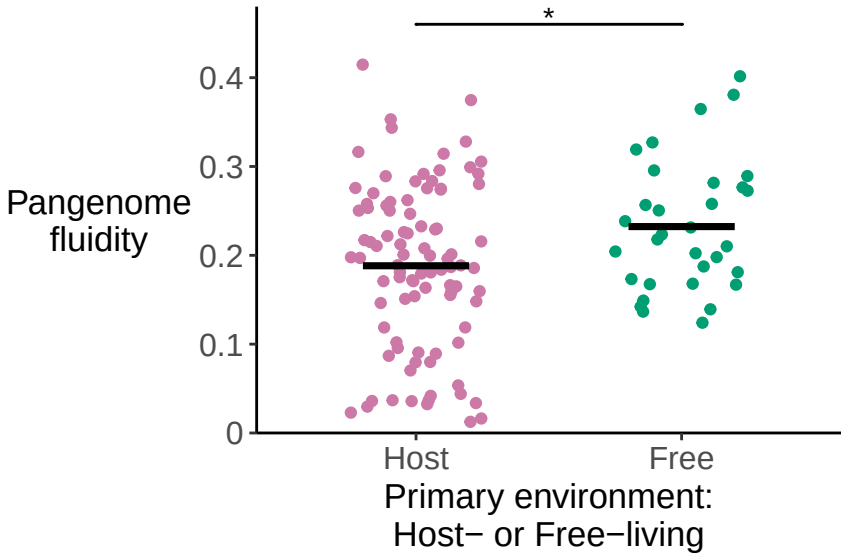


Figure S2: Whether a species' primary environment is host-associated or free-living correlates with pangenome fluidity.

1.3 Host, Mostly-Host, Mostly Free and Free

Finally, we combined information from our first and second method of categorising species into a third variable. To do this, we split the 'Both' category from the first method into two, based on whether species' primary environment was host-associated or free-living. This meant we had four categories: 'Host' (species that are always host-associated), 'Mostly host' (species with a host-associated primary environment, but which sometimes are free-living), 'Mostly free' (species with a free-living primary environment, but which are sometimes host-associated), and 'Free' (species which are always host-associated). We used this method of categorising species in the main text of the paper.

1.3.1 Linear model

We found that species which are always host-associated have a lower pangenome fluidity than species which are at least sometimes free-living.

Table S6: Results from a linear model with pangenome fluidity as the response variable and whether a species' is always host-associated, Mostly host-associated, mostly free-living, or always free-living as the explanatory variable.

	Estimate	Std.Error	t.value	p.value	signif.
(Intercept)	0.15546	0.01031	15.074	0.0000	***
Host_free_4Mostly_host	0.09240	0.01731	5.337	0.0000	***
Host_free_4Mostly_free	0.06983	0.01828	3.820	0.0002	***
Host_free_4Free	0.12632	0.04125	3.062	0.0027	**

1.3.2 MCMCglmm

We also found analogous results when we ran a MCMCglmm to control for phylogenetic relationships between species.

Table S7: Results from a MCMCglmm with pangenome fluidity as the response variable, whether a species' is always host-associated, mostly host-associated, mostly free-living or always free-living as the fixed effect, and phylogeny as a random effect.

	post.mean	l-95% CI	u-95% CI	eff.samp	pMCMC	signif.
(Intercept)	0.17864	0.10320	0.2461	4368	0.0002128	***
Host_free_4Mostly_host	0.08650	0.05166	0.1207	4700	0.0002128	***
Host_free_4Mostly_free	0.06499	0.02969	0.1021	4198	0.0004255	***
Host_free_4Free	0.11864	0.04084	0.2016	4700	0.0051064	**

	R-squared value
Fixed effect	0.1465
Random effect	0.4655
Total model	0.6120

We can view these results with the accompanying data in the figure below: this is the same as Figure 2a in the main text.

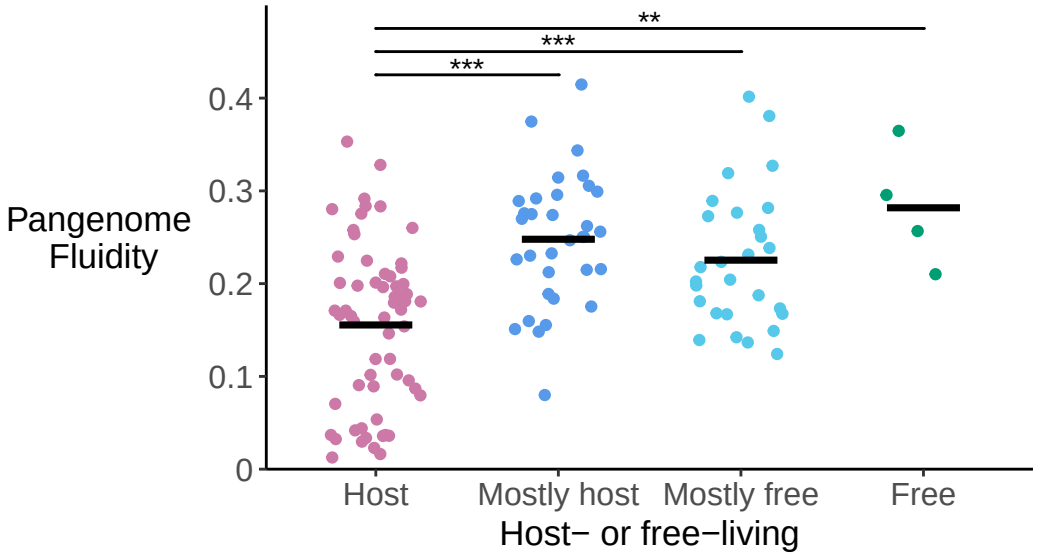


Figure S3: Whether a species' is host-associated, free-living, or a mixture correlates with pangenome fluidity.

1.4 Summary

Taken together, we find that whether a species is free-living or host-associated appears to be a key correlate of pangenome fluidity across bacterial species. Host-associated species have a lower pangenome fluidity than free-living species, while species that are sometimes host-associated and sometimes free-living have an intermediate pangenome fluidity.

2 Within hosts, which of four key lifestyle traits matter for shaping pangenome fluidity?

However, categorising species as free-living or host-associated does not account for variation within each of the categories. All but four of our species have been found to live, at least sometimes, associated with hosts,

and there is clearly much variation in pangenome fluidity left unexplained. For example, some species live across multiple hosts, while others only ever live inside one; some species can live in multiple sites inside hosts, while others are restricted to certain areas of the host, or even only within their cells.

Therefore, there are likely to be many lifestyle traits which differentiate host-associated bacteria. We can predict how each trait might influence a species' pangenome fluidity. We can then examine whether pangenome fludity is correlated with such traits in the direction we would expect, to resolve whether they are acting directly to shape pangenome fluidity across species.

To do this, we categorised our species into four additional lifestyle traits that vary among host-associated species, each with a predicted impact on pangenome fluidity. These traits were: (i) nature of host association (obligate or facultative); (ii) location within host (intracellular or extracellular); (iii) effect on host (pathogen or mutualist); motility (non-motile or motile). For each, the first category listed is predicted to have a lower pangenome fluidity than the second category. For several of these traits, we included an additional 'Both' category to capture species which can live both inside and outside cells, act both as a pathogen and a mutualist, or exist as both non-motile and motile cells.

First, we found that all four traits correlated with pangenome fluidity in the direction we expected.

2.1 Nature of host-association: obliagte or facultative

We found that obligately host-associated species had a lower pangenome fluidity than facultatively host-associated species. There were 115 species for which we had data on their nature of host association. First, we found this result using a simple linear model.

Table S8: Results from a linear model with pangenome fluidity as the response variable and whether a species' is obligately or facultatively host-associated as the explanatory variable.

	Estimate	Std.Error	t.value	p.value	signif.
(Intercept)	0.1197	0.01487	8.051	0	***
Obligate_facultativeFacultative	0.1028	0.01719	5.978	0	***

We also found the same when we controlled for phylogeny using a MCMCglmm. Whether a species was obligately or facultatively host-associated explained around 14% of the variance in pangenome fluidity across species.

Table S9: Results from a MCMCglmm with pangenome fluidity as the response variable, whether a species' is obligately or facultatively host-associated as the fixed effect, and phylogeny as a random effect.

	post.mean	l-95% CI	u-95% CI	eff.samp	pMCMC	signif.
(Intercept)	0.13360	0.06613	0.2127	4700	0.0029787	**
Obligate_facultativeFacultative	0.09051	0.05305	0.1257	4158	0.0002128	***

	R-squared value
Fixed effect	0.1409
Random effect	0.4156
Total model	0.5565

We can view these results from the MCMCglmm alongside the data in the following figure.

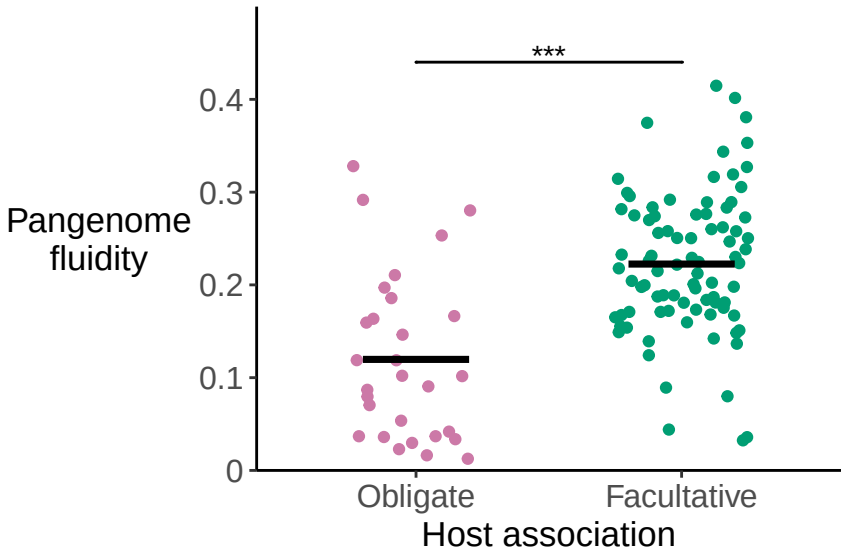


Figure S4: Obligately host-associated species have a lower pangenome fluidity than facultatively host-associated species

2.2 Location within host: intracellular or extracellular

We found that species which lived inside their host(s) cells had a lower pangenome fluidity than species which lived outside host cells. There were 120 species for which we had data on their location within hosts. We found these results first using a simple linear model.

Table S10: Results from a linear model with pangenome fluidity as the response variable and whether a species lives inside or outside host cells as the explanatory variable.

	Estimate	Std.Error	t.value	p.value	signif.
(Intercept)	0.145335	0.01701	8.54215	0.0000	***
Intra_or_extracellularBoth	-0.001501	0.02190	-0.06855	0.9455	
Intra_or_extracellularExtracellular	0.092641	0.01950	4.75111	0.0000	***

We then controlled for phylogeny with a MCMCglmm and found analogous results. Location within host accounted for 13% of the variance in pangenome fluidity across species.

Table S11: Results from a MCMCglmm with pangenome fluidity as the response variable, whether a species lives inside or outside host cells as the fixed effect, and phylogeny as a random effect.

	post.mean	l-95% CI	u-95% CI	eff.samp	pMCMC	signif.
(Intercept)	0.17400	0.099456	0.24482	3865	0.0002128	***
Intra_or_extracellularBoth	-0.02836	-0.078041	0.02051	3132	0.2638298	
Intra_or_extracellularExtracellular	0.05400	0.003691	0.10867	2925	0.0489362	*

	R-squared value
Fixed effect	0.1288
Random effect	0.3842
Total model	0.5130

We can also set species which are extracellular as the intercept, to examine if species which are extracellular have significantly higher pangenome fluidity than species in the intermediate 'both' category.

Table S12: Results from a MCMCglmm with pangenome fluidity as the response variable, whether a species lives inside or outside host cells as the fixed effect, and phylogeny as a random effect; instead with extracellular as the intercept.

	post.mean	l-95% CI	u-95% CI	eff.samp	pMCMC	signif.
(Intercept)	0.17319	0.100877	0.24595	4334	0.0002128	***
Intra_or_extracellularBoth	-0.02754	-0.079989	0.02001	3174	0.2782979	
Intra_or_extracellularExtracellular	0.05529	0.004277	0.10757	2747	0.0378723	*

	R-squared value
Fixed effect	0.1288
Random effect	0.3842
Total model	0.5130

We can view these results together with the data in the following figure - this is the same as Figure 2c in the main text.

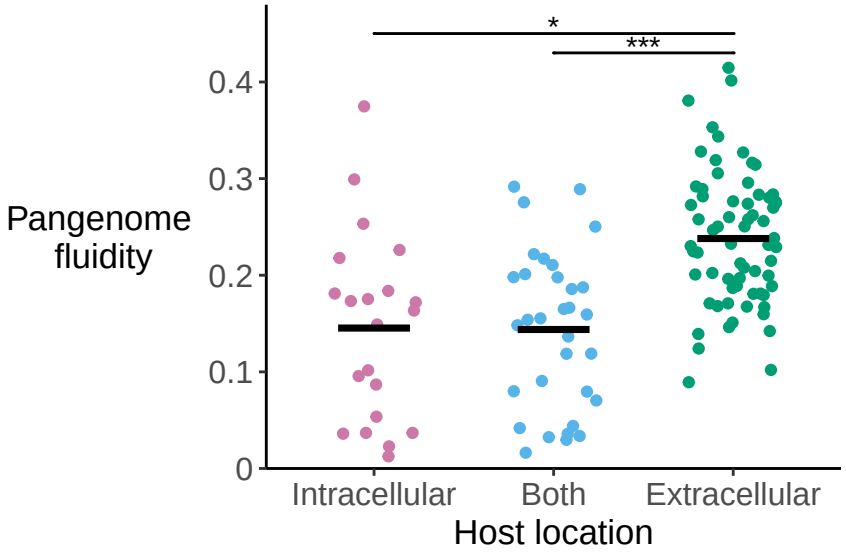


Figure S5: Intracellular species have a lower pangenome fluidity than extracellular species

2.3 Effect on host: pathogen or mutualist

We found that pathogenic species had lower pangenome fluidity compared to mutualist species and also compared to species which were sometimes pathogenic and sometimes mutualist. This was across 119 species for which we had effect on host data. We first found these results with a simple linear model.

Table S13: Results from a linear model with pangenome fluidity as the response variable and whether a species' effect on its host as the explanatory variable.

	Estimate	Std.Error	t.value	p.value	signif.
(Intercept)	0.14736	0.01112	13.250	0.00e+00	***
Effect_on_hostBoth	0.08559	0.01633	5.242	7.00e-07	***
Effect_on_hostMutualist	0.09109	0.02040	4.466	1.86e-05	***

We then found analogous results when controlling for phylogeny using a MCMCglmm.

Table S14: Results from a MCMCglmm with pangenome fluidity as the response variable, whether a species has a pathogenic or mutualistic effect on its host(s) as the fixed effect, and phylogeny as a random effect.

	post.mean	l-95% CI	u-95% CI	eff.samp	pMCMC	signif.
(Intercept)	0.15350	0.08229	0.22187	4700	0.0002128	***
Effect_on_hostBoth	0.06397	0.03243	0.09642	4408	0.0004255	***
Effect_on_hostMutualist	0.10048	0.05875	0.14133	4700	0.0002128	***

	R-squared value
Fixed effect	0.1472
Random effect	0.4243
Total model	0.5715

We can view these results from the MCMCglmm together with the data in the following graph, which is the same as Figure 2d in the main text.

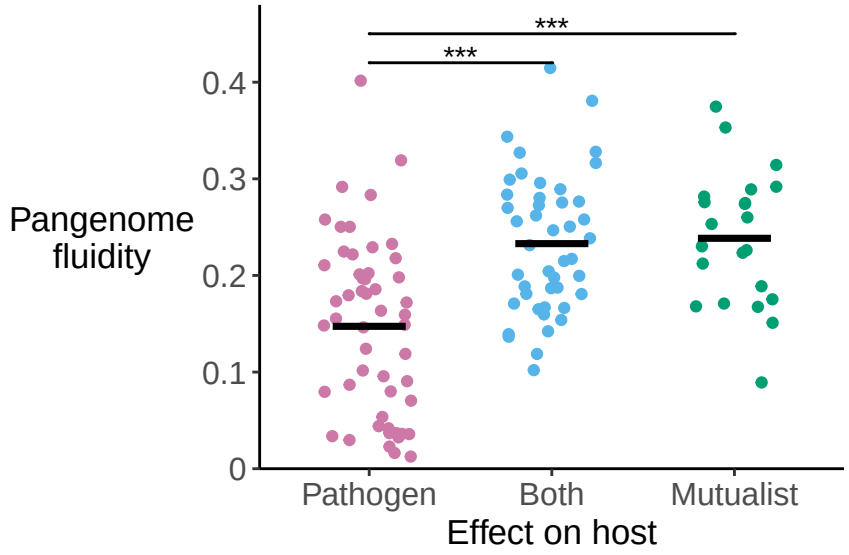


Figure S6: Non-motile species have a lower pangenome fluidity than motile species

2.4 Motility: non-motile or motile

We found that non-motile species had lower pangenome fluidity than motile species. This was across 126 species, since we had motility information for all species. First, we found these results in a simple linear model.

Table S15: Results from a linear model with pangenome fluidity as the response variable and a species' motility as the explanatory variable.

	Estimate	Std.Error	t.value	p.value	signif.
(Intercept)	0.181224	0.01064	17.033	0.0000	***
MotilityBoth	-0.007697	0.03737	-0.206	0.8371	
MotilityMotile	0.045212	0.01616	2.797	0.0060	**

We also found analogous results when controlling for phylogeny with a MCMCglmm.

Table S16: Results from a MCMCglmm with pangenome fluidity as the response variable, a species' motility as the fixed effect, and phylogeny as a random effect.

	post.mean	l-95% CI	u-95% CI	eff.samp	pMCMC	signif.
(Intercept)	0.198631	0.12429	0.27720	4486	0.0002128	***
MotilityBoth	0.009009	-0.05522	0.07388	4700	0.7800000	
MotilityMotile	0.049866	0.01350	0.08229	4817	0.0051064	**

	R-squared value
Fixed effect	0.04646
Random effect	0.55028
Total model	0.59674

We can view these MCMCglmm results together with the data in the following graph, which is the same as Figure 2e in the main text.

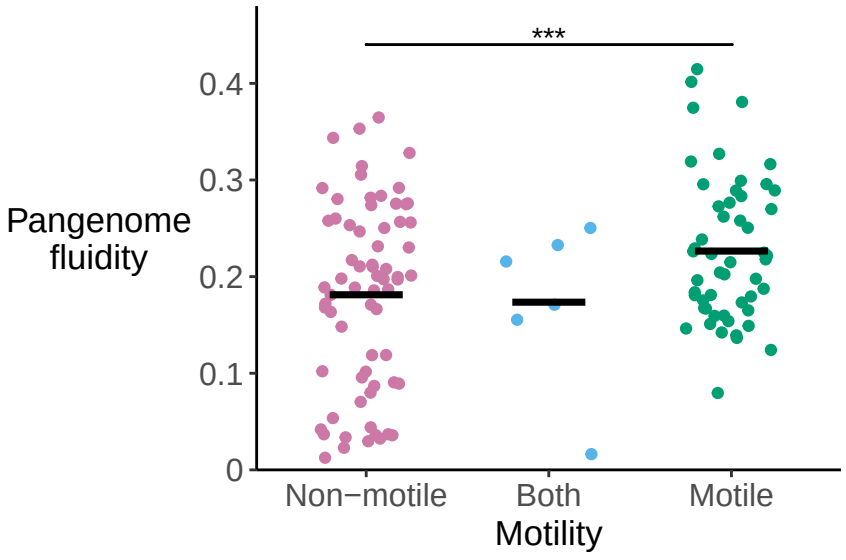


Figure S7: Non-motile species have a lower pangenome fluidity than motile species

2.5 Summary

We found that all four lifestyle traits which might vary across host-associated species were significantly correlated with pangenome fludity in the direction we predicted. Species have lower pangenome fluidity if

they are obligate compared to facultative, intracellular compared to extracellular, pathogens compared to mutualists and non-motile compared to motile.

2.6 Correlations between lifestyle traits

Next, we wanted to explore the extent to which each of these lifestyle traits directly influenced species' pangenome fluidity, and how much of this influence was independent of the other lifestyle traits.

First, we explored how much variance in pangenome fluidity was explained by the four lifestyle traits together, taking into account phylogeny. We did this across the 115 species for which we had data on all four of the lifestyle traits.

Table S17: Results from a MCMCglmm with pangenome fluidity as the response variable, four lifestyle traits as fixed effects, and phylogeny as a random effect.

	post.mean	l-95% CI	u-95% CI	eff.samp	pMCMC	signif.
(Intercept)	0.12852	0.053273	0.20370	4470	0.001277	**
Obligate_facultativeFacultative	0.05012	0.007215	0.08961	4700	0.014894	*
Intra_or_extracellularBoth	-0.02631	-0.074652	0.02284	4314	0.285532	
Intra_or_extracellularExtracellular	0.01899	-0.035994	0.06971	3757	0.475319	
Effect_on_hostBoth	0.03233	-0.003372	0.06834	4700	0.083404	.
Effect_on_hostMutualist	0.05804	0.013520	0.10102	4700	0.010638	*
MotilityBoth	-0.01068	-0.076686	0.05867	4700	0.749787	
MotilityMotile	0.01643	-0.017269	0.05234	4700	0.363404	

	R-squared value
Fixed effect	0.2519
Random effect	0.3594
Total model	0.6113

We found that overall, the four traits together explained ~25.7% of the total variance in pangenome fluidity across 115 species.

This is a high R-squared value for biological data. However, the sum of the R-squared values from the four models where each lifestyle trait was a single fixed effect is 0.458, or 45.8% of the variance explained. Summing the values in this way assumes that the variances in pangenome fluidity explained by each trait are independent from the variances explained by the other traits. However, we instead find that all four traits explain only 25.7% of the variance in pangenome fluidity, just over half the summed total. This suggests that the influence of each lifestyle trait on pangenome fluidity is not independent from one another. This could be because one or more of the lifestyle traits might be themselves correlated. For example, whether a species is obligate or facultative may correlate with whether a species lives inside or outside cells and its motility.

Next, we used several complimentary methods to explore potential correlations between these four lifestyle traits.

2.6.1 Phylogenetic regressions

First, we used phylogenetic regression models to assess evidence for correlations between all pairs of the four lifestyle traits across 115 species.

For each trait we assigned a 0 to species with the less variable lifestyle (obligate, intracellular, pathogenic, non-motile) and a 1 to species with the more variable lifestyle (facultative, extracellular, mutualistic, motile). For species which were in an intermediate category for the trait ('both'), we assigned them a value of 0.5.

This meant host association was a binary trait (obligate=0,facultative=1), and host location, effect on host and motility were treated as discrete variables with three evenly spaced levels.

When host association was the response variable we used phylogenetic logistic regression models (`phyloglm()` function in the `phylolm` R package). For models where the other three traits were the response variable, we used phylogenetic linear regressions (`phylolm()` function in `phylolm` package in R).

Table S18: Phylogenetic logistic regression correlated evolution model between host association and host location

	Estimate	Std.Error	t.value	p.value	signif.
(Intercept)	-0.01722	0.4806	-0.03583	0.9714	
IE	1.42135	0.6457	2.20139	0.0277	*

Table S19: Results from a phylogenetic logistic regression correlated evolution model between host association and effect on host

	Estimate	Std.Error	t.value	p.value	signif.
(Intercept)	0.3358	0.3651	0.9197	0.3577	
EH	1.6266	0.6793	2.3945	0.0166	*

Table S20: Results from a phylogenetic logistic regression correlated evolution model between host association and motility

	Estimate	Std.Error	t.value	p.value	signif.
(Intercept)	0.3323	0.3776	0.8799	0.3789	
M	1.4358	0.5132	2.7979	0.0051	**

Table S21: Results from a phylogenetic logistic regression correlated evolution model between host location and effect on host

	Estimate	Std.Error	t.value	p.value	signif.
(Intercept)	0.5516	1.02231	0.5396	0.5905	
EH	0.2583	0.06062	4.2609	0.0000	***

Table S22: Results from a phylogenetic logistic regression correlated evolution model between host location and motility

	Estimate	Std.Error	t.value	p.value	signif.
(Intercept)	0.61832	1.09915	0.5625	0.5749	
M	0.02721	0.04064	0.6696	0.5045	

Table S23: Results from a phylogenetic logistic regression correlated evolution model between effect on host and motility

	Estimate	Std.Error	t.value	p.value	signif.
(Intercept)	0.27634	1.57893	0.175	0.8614	
M	0.06075	0.05837	1.041	0.3003	

The results of these phylogenetic regressions suggest that species which were obligately associated with hosts were also more likely to live inside host cells, act as pathogens, and be non-motile, while species which lived inside cells were more likely to act as pathogens (Host association vs: (i) Host location, p=0.028; (i) Effect on host, p=0.017; (iii) Motility, p=0.005; Host location vs. effect on host: p<0.001).

2.6.2 Pagel's method of correlated evolution

Next, we used Pagel's 1994 method of correlated evolution to test for correlated evolution between pairs of binary traits: Pagel, M. (1994) Detecting correlated evolution on phylogenies: A general method for the comparative analysis for discrete characters. Proceedings of the Royal Society B, 255, 37-45. We implemented this method using the `fitPagel()` function of the R package `phytools`.

This method defines four potential states based on two binary traits:(0,0), (1,0), (0,1), and (1,1). The distribution of these states are then examined across the phylogeny to identify whether the each of the traits is evolving between state 0 and 1 independently, or instead are in some way dependent on the other trait.

To use this method, we needed to convert our traits into binary traits. To do this, we simplified the traits into binary variables, where 0 was the lower variability lifestyle (obligate, intracellular, pathogenic and non-motile) and 1 was the higher variability lifestyle (facultative, extracellular, mutualistic and motile), as before. We needed to merge species from the 'both' category to either a 0 or 1 for several traits. We did this by: (i) merging species which were sometimes intracellular together with those that were always intracellular; (ii) merging species which were sometimes pathogenic with species that were always pathogenic; (iii) merging species which were sometimes motile with species that were always motile.

We found that the evolution of species' host association was significantly correlated with the evolution of species' host location.

Table S24: Results from a Pagel's correlated evolution model between host association and host location

			Log-likelihood	AIC
Likelihood ratio	13.0450	Independent	-95.7826	-89.2601
p-value	0.0111	Dependent	199.5652	194.5202

We found that the evolution of species' host association was independent from the evolution of species' effect on host.

Table S25: Results from a Pagel's correlated evolution model between host association and effect on host

			Log-likelihood	AIC
Likelihood ratio	8.4754	Independent	-85.9946	-81.7569
p-value	0.0756	Dependent	179.9891	179.5138

We found that the evolution of species' host association was significantly correlated with the evolution of species' motility.

Table S26: Results from a Pagel's correlated evolution model between host association and motility

			Log-likelihood	AIC
Likelihood ratio	21.7542	Independent	-116.2164	-105.3393
p-value	0.0002	Dependent	240.4328	226.6785

We found the evolution of species' host location was independent from the evolution of species' effect on host.

Table S27: Results from a Pagel's correlated evolution model between host location and effect on host

			Log-likelihood	AIC
Likelihood ratio	1.8717	Independent	-79.0873	-78.1515
p-value	0.7593	Dependent	166.1747	172.3030

We found the evolution of species' host location was independent from the evolution of species' motility.

Table S28: Results from a Pagel's correlated evolution model between host location and motility

			Log-likelihood	AIC
Likelihood ratio	4.3251	Independent	-109.3092	-107.1466
p-value	0.3638	Dependent	226.6183	230.2932

We found the evolution of species' effect on host was independent from the evolution of species' motility.

Table S29: Results from a Pagel's correlated evolution model between effect on host and motility

			Log-likelihood	AIC
Likelihood ratio	6.2765	Independent	-99.5211	-96.3829
p-value	0.1794	Dependent	207.0422	208.7657

We found that obligate host association was correlated with the evolution of living inside cells and with reduced motility (Pagel's correlated evolution model; Host association vs: (i) Host location, $p=0.011$; (ii) Motility, $p<0.001$). This agreed with the results from our phylogenetic regressions.

However, in contrast, our results using Pagel's method do not suggest that the evolution of host association and effect on host, and also the evolution of host location and effect on host, are correlated. This might be due to host location and effect on host now being treated as a binary variable, removing the intermediate category, and thus some of the variation across species.

2.6.3 Binomial generalised linear models

Next, using the same binary variables, we ran binomial generalised linear models (glms) between the pairs of the four traits. We did this to examine whether species with a more variable lifestyle for one trait were more likely to also have a more variable lifestyle for the other trait. This asks simply whether species are more likely to have one lifestyle if they have another, independent of phylogeny.

Table S30: Results from a binomial generalised linear model with species' host association as the response variable and species' host location as the explanatory variable.

	Estimate	Std.Error	t.value	p.value	signif.
(Intercept)	0.04082	0.2858	0.1428	0.8864	
IE	2.46061	0.5459	4.5070	0.0000	***

Table S31: Results from a binomial generalised linear model with species' host association as the response variable and species' effect on host as the explanatory variable.

	Estimate	Std.Error	t.value	p.value	signif.
(Intercept)	0.8422	0.2261	3.726	0.0002	***
EH	2.2023	1.0481	2.101	0.0356	*

Table S32: Results from a binomial generalised linear model with species' host association as the response variable and species' motility as the explanatory variable.

	Estimate	Std.Error	t.value	p.value	signif.
(Intercept)	0.3075	0.2635	1.167	0.2432	
M	2.2575	0.5819	3.879	0.0001	***

Table S33: Results from a binomial generalised linear model with species' host location as the response variable and species' effect on host as the explanatory variable.

	Estimate	Std.Error	t.value	p.value	signif.
(Intercept)	0.3075	0.2635	1.167	0.2432	
M	2.2575	0.5819	3.879	0.0001	***

Table S34: Results from a binomial generalised linear model with species' host location as the response variable and species' motility as the explanatory variable.

	Estimate	Std.Error	t.value	p.value	signif.
(Intercept)	0.0339	0.2604	0.1302	0.8964	
M	0.5539	0.3816	1.4516	0.1466	

Table S35: Results from a binomial generalised linear model with species' effect on host as the response variable and species' motility as the explanatory variable.

	Estimate	Std.Error	t.value	p.value	signif.
(Intercept)	-1.1676	0.3060	-3.815	0.0001	***
M	-0.6242	0.4894	-1.275	0.2022	

The results of these binomial glms suggest that host association is significantly correlated with the three other traits, and host location is significantly correlated with a species' effect on host. This is qualitatively the same as the results from the phylogenetic regressions of the traits when including the intermediate categories, although those models accounted for phylogeny and these models do not.

We can visualise the correlations between traits in the figure below. For example, in the first panel, we can see that intracellular species appear equally likely to be obligate or facultative, while extracellular species are much more likely to be facultative than obligate.

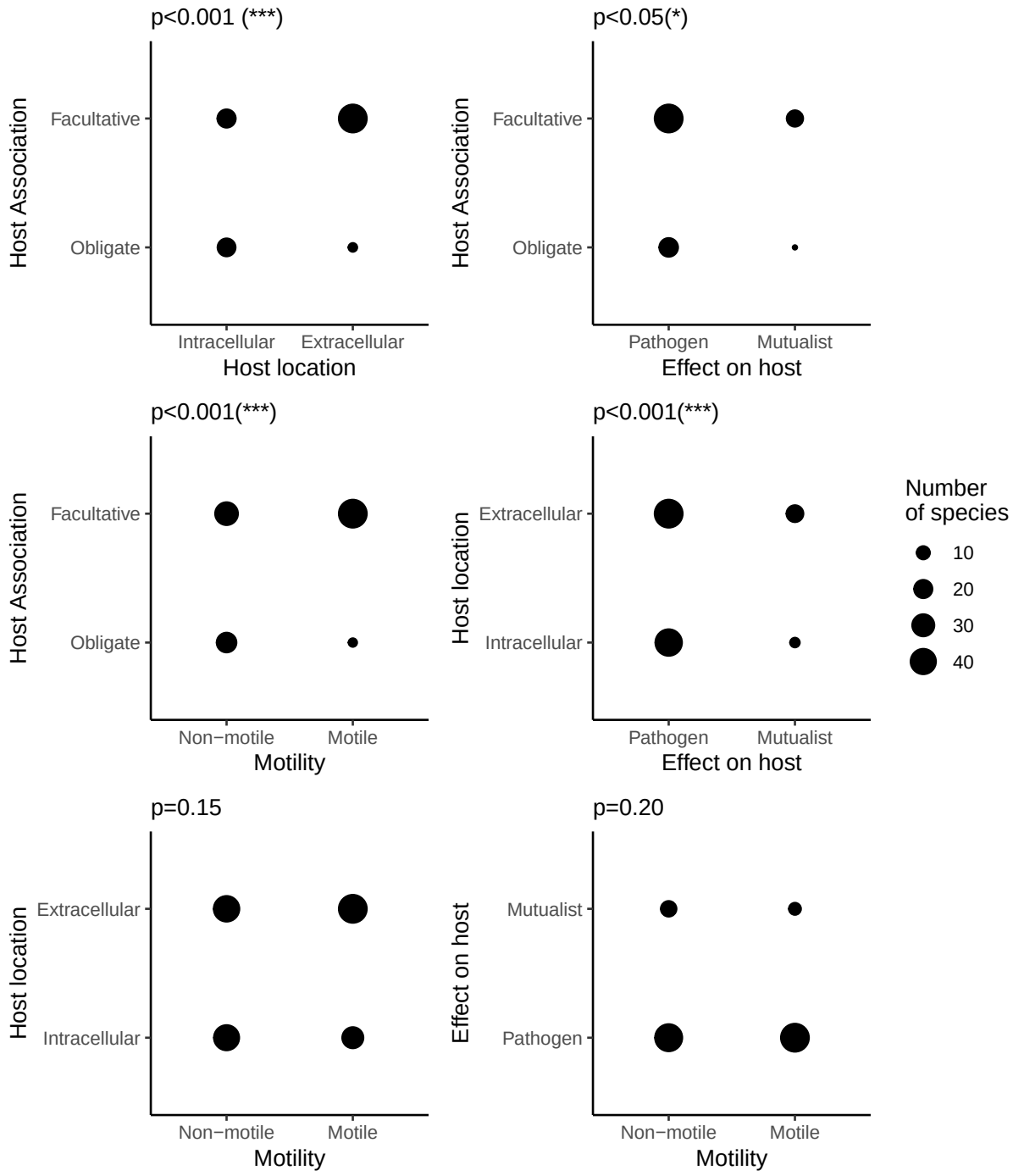


Figure S8: Associations between pairs of four lifestyle traits.The size of circles correspond to the number of species in one of four combinations of lifestyles for each trait.

However, this approach does not take into account phylogenetic history and relationships between species, instead assuming all species are independent data points.

2.6.4 Phylogenetic logistic regressions

To control for phylogenetic relationships, we ran phylogenetic logistic regression models as described in: Ives, A. R. and T. Garland, Jr. 2010. "Phylogenetic logistic regression for binary dependent variables". Systematic Biology 59:9-26. We used the `phyloglm()` function within the `phylolm` package to do this, with the `method` set to "logistic_MPLE".

We used this method to examine pairs of traits in binary form, as described earlier.

Table S36: Results from a phylogenetic logistic regression model with species' host association as the response variable and species' host location as the explanatory variable.

	Estimate	Std.Error	t.value	p.value	signif.
(Intercept)	0.2621	0.3899	0.6723	0.5014	
IE	1.2196	0.5439	2.2424	0.0249	*

Table S37: Results from a phylogenetic logistic regression model with species' host association as the response variable and species' effect on host as the explanatory variable.

	Estimate	Std.Error	t.value	p.value	signif.
(Intercept)	0.6855	0.3451	1.986	0.0470	*
EH	1.0807	0.7521	1.437	0.1508	

Table S38: Results from a phylogenetic logistic regression model with species' host association as the response variable and species' motility as the explanatory variable.

	Estimate	Std.Error	t.value	p.value	signif.
(Intercept)	0.3703	2.06755	0.1791	0.8582	
M	0.2805	0.07109	3.9460	0.0001	***

Table S39: Results from a phylogenetic logistic regression model with species' host location as the response variable and species' effect on host as the explanatory variable.

	Estimate	Std.Error	t.value	p.value	signif.
(Intercept)	0.06456	0.3804	0.1697	0.8652	
EH	0.38864	0.4685	0.8296	0.4068	

Table S40: Results from a phylogenetic logistic regression model with species' host location as the response variable and species' motility as the explanatory variable.

	Estimate	Std.Error	t.value	p.value	signif.
(Intercept)	0.05666	0.3788	0.1496	0.8811	
M	0.08858	0.2377	0.3727	0.7094	

Table S41: Results from a phylogenetic logistic regression model with species' effect on host as the response variable and species' motility as the explanatory variable.

	Estimate	Std.Error	t.value	p.value	signif.
(Intercept)	-1.5006	0.4477	-3.3517	0.0008	***
M	-0.2576	0.4506	-0.5717	0.5675	

Across the phylogeny of 115 species, the phylogenetic logistic regressions suggested that host association was significantly correlated with both host location and motility. Both were positive correlations, meaning that species which were obligate were more likely to be intracellular and non-motile, while facultative species were more likely to be extracellular and motile.

However, host association and host location were no longer correlated with effect on host. This suggests that these correlations can be explained by co-ancestry of lifestyle traits, at least when the traits are considered as binary, and the intermediate category removed.

2.6.5 Summary

Overall, we used several complimentary methods to assess whether our four lifestyle traits were correlated with one another.

We found strong evidence that species' host association was correlated with species' host location and species' motility. These two pairs of traits were significantly correlated in all tests we performed, and regardless of whether we included an intermediate category for host location and motility. Species which were obligate, compared to facultative, were more likely to also be intracellular and non-motile, compared to extracelluar and motile. We also found evidence that the distribution of these traits across the phylogeny was not independent, potentially suggesting the host association evolves together with host location and motility.

We also found some evidence that species' host association and host location were correlated with species' effect on host. We found significant correlations between these two pairs both when asking simply if the traits were correlated, and also when using phylogenetic regressions to account for similarity beetween species due to co-ancestry. However, several methods we used required us to convert the traits into binary variables, meaning we needed to merge species which were only sometimes pathogenic with species which were always pathogenic, and merge species which were sometimes motile with those that were always motile. When we used Pagel's method of correlated evolution and phylogenetic logistic regression models to assess correlations between these binary traits, these two pairs of traits were no longer significantly correlated.

These correlations across the different lifestyle traits make it hard to determine the underlying causality for pangenome fluidity. Each of the lifestyle traits could lead to less fluid pangenomes directly, or instead only correlate with pangenome fluidity because they are influenced by another trait which directly influences pangenome fluidity. Alternatively, a trait could indirectly influence pangenome fluidity by influencing another lifestyle trait, which then influences pangenome fluidity directly. Phylogenetic correlations alone are unable to distinguish between a huge number of possibilities.

2.7 Phylogenetic path analysis

We resolved this problem with phylogenetic path analysis. This method is based upon the theory of causal inference, which suggests that while correlation does not equal causation, correlation, if not due to chance, always implies an underlying causal structure. Using a path analysis, one can compare support for multiple hypothesised causal models by constructing a set of phylogenetic linear models which must be supported in order for the model to not be rejected. For example, if A caused both B and C, a linear model could be constructed in which B would no longer correlate with C once A was taken into account. If many variables are included, linear models can be constructed for each causal pathway included in a potential model of causation. Support for any models not rejected can then be compared.

We used this method to infer whether each trait had a direct causal influence on species' pangenome fluidity, or whether instead one or a few traits were the main drivers.

We coded the lifestyle traits as described previously in section X.X, where 0 corresponds to the lower variability lifestyle (obligate, intracellular, pathogenic, non-motile), 1 corresponds to the higher variability lifestyle (facultative, extracellular, mutualistic, motile), and 0.5 corresponds to species which have an intermediate lifestyle.

We used the R package `phylopath` for our phylogenetic path analyses.

For a helpful guide on how to get started with phylogenetic path analyses in R, please see: https://ax3man.github.io/phylopath/articles/intro_to_phylopath.html. Please also see Chapter 8 of this book: L. Z. Garamszegi, Modern Phylogenetic Comparative Methods and Their 201 Application in Evolutionary Biology, which provides a great background on the theory of causal inference and its application to phylogenetic comparative methods.

2.7.1 Simple models with no paths between lifestyle traits

First we compared a set of simple causal models, which varied by which of the four lifestyle traits had a direct causal influence on pangenome fluidity. These models can be viewed in the following figure. An arrow means that the trait at the beginning of the arrow causes the trait at the end of the arrow.

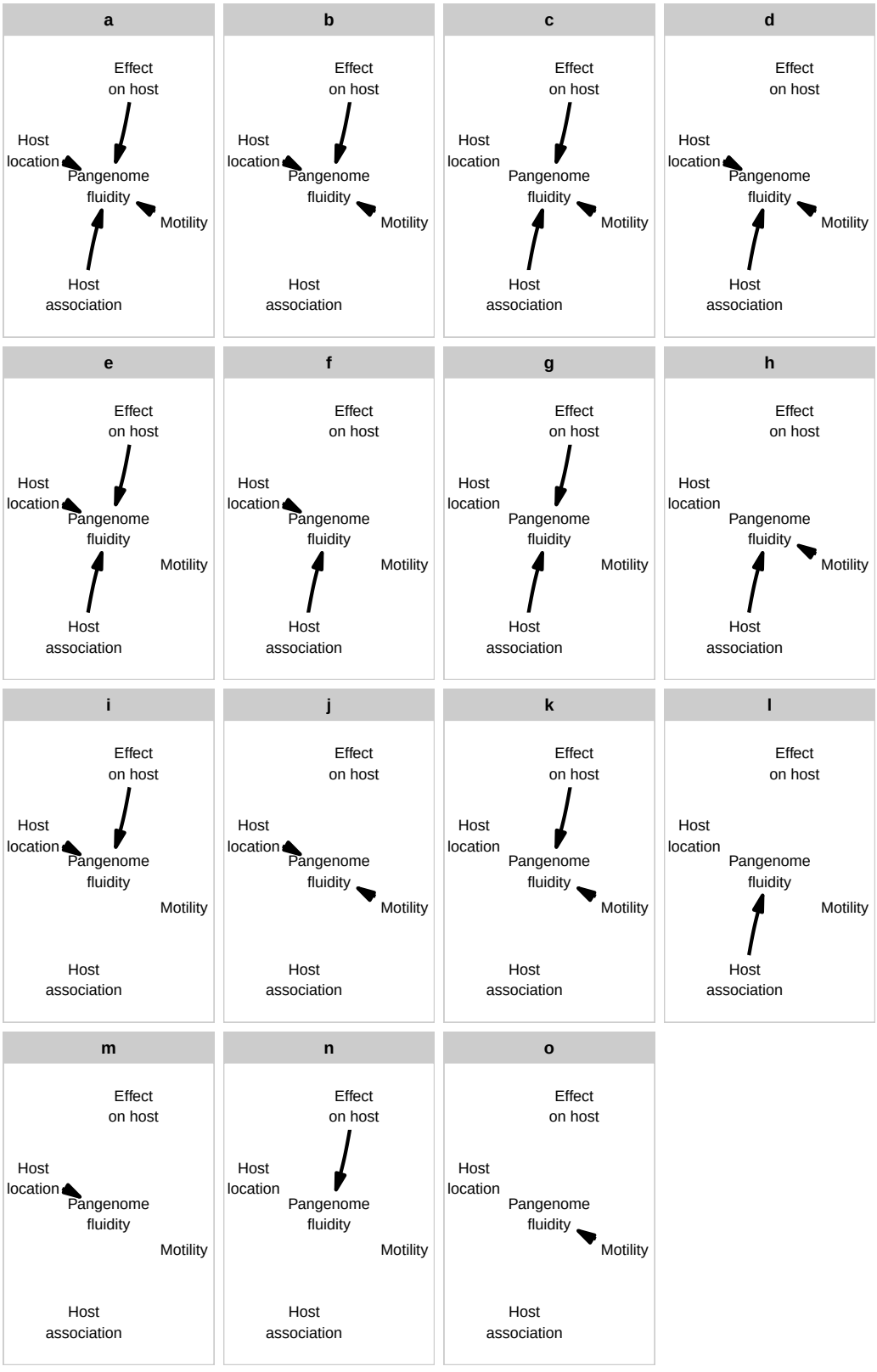


Figure S9: Set of simple models, varying by which lifestyle traits cause pangenome fluidity

We then used the `phylo_path()` function to compare support for this set of simple models. We used the default model of evolution, Pagel's lambda, and the default method for the evolution of binary traits (which applies to host assocation only), `"logistic_MPLE"` (maximizes the penalized likelihood of the logistic regression).

We found that all of these simple models were rejected as a plausible model of causation: all had a p-value of <0.001.

A path analysis works by constructing a series of linear models based on the "conditional independencies" assumed by each model's structure. If a model shows that A causes both B and C, both B and C are said to be conditionally independent from one another. This can be expressed as linear model, where the slope of B~A+C should not be significantly different from zero, because B should no longer correlate with C once the causal influence of A is controlled for. This linear model is a conditional independency; a model examined by a path analysis may have multiple of these, and for each the p-value comparing the slope to zero must be greater than 0.05 for the model to have good support.

For each of the simple models we examined to be rejected, there must be at least one conditional independency with a p-value of less than 0.05. By having no causal links between the lifestyle traits, we are assuming that all of the traits are conditionally independent of one another: put more simply, that the traits are not correlated with each other once their independent influence on pangenome fluidity is accounted for. However, we found strong evidence that several of the lifestyle traits are themselves correlated in the previous section.

Therefore, these simple models of causation are rejected because the lifestyle traits themselves are correlated, and their evolution potentially depends on the evolution of another trait(s).

Instead, we need to expand our model choice to allow for causal links between the lifestyle traits themselves. We did this by examining the conditional independencies which were significant in the simple models, adding causal links between those traits.

2.7.2 Complex models with paths between lifestyle traits

We then compared this more complex set of models, which can be seen in the following figure.

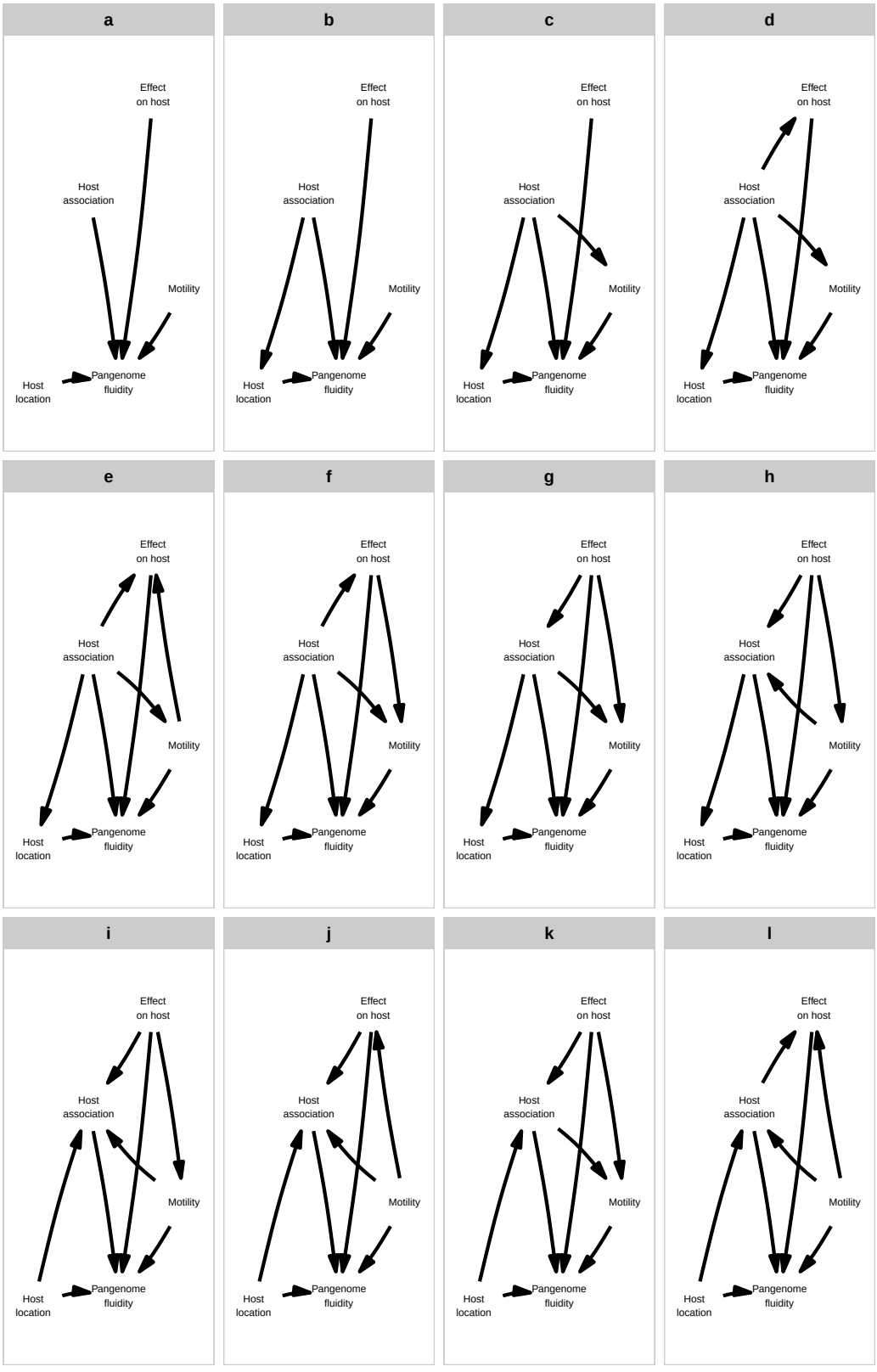


Figure S10: Set of more complex models, varying by which lifestyle traits cause pangenome fluidity and how they might cause each other.

We again used a path analysis to compare support for these more complex models. We now found that there were a number of models which had a p-value of >0.05, meaning they cannot be rejected as a plausible model of causation bewteen the variables.

Table S42: Comparison of set of more complex causal models, varying by which lifestyle trait(s) directly influence pangenome fluidity and how they might cause each other.

	w	p	CICc
j	0.3186	0.5314	32.7643
i	0.2725	0.4821	33.0764
l	0.0760	0.1971	35.6311
g	0.0671	0.1796	35.8784
h	0.0671	0.1796	35.8784
f	0.0671	0.1796	35.8784
e	0.0626	0.1703	36.0180
d	0.0552	0.1620	36.2704
k	0.0105	0.0407	39.5873
c	0.0032	0.0262	41.9661
b	0.0000	0.0000	71.6239
a	0.0000	0.0000	77.5508

We can then view support for the models by comparing their value of w which is a measure of model support based on the C statistic Information Criteria (CIC), which is itself a modified version of Akaike Information Criteria (AIC) for comparing model support. A higher w value means the model has higher support.

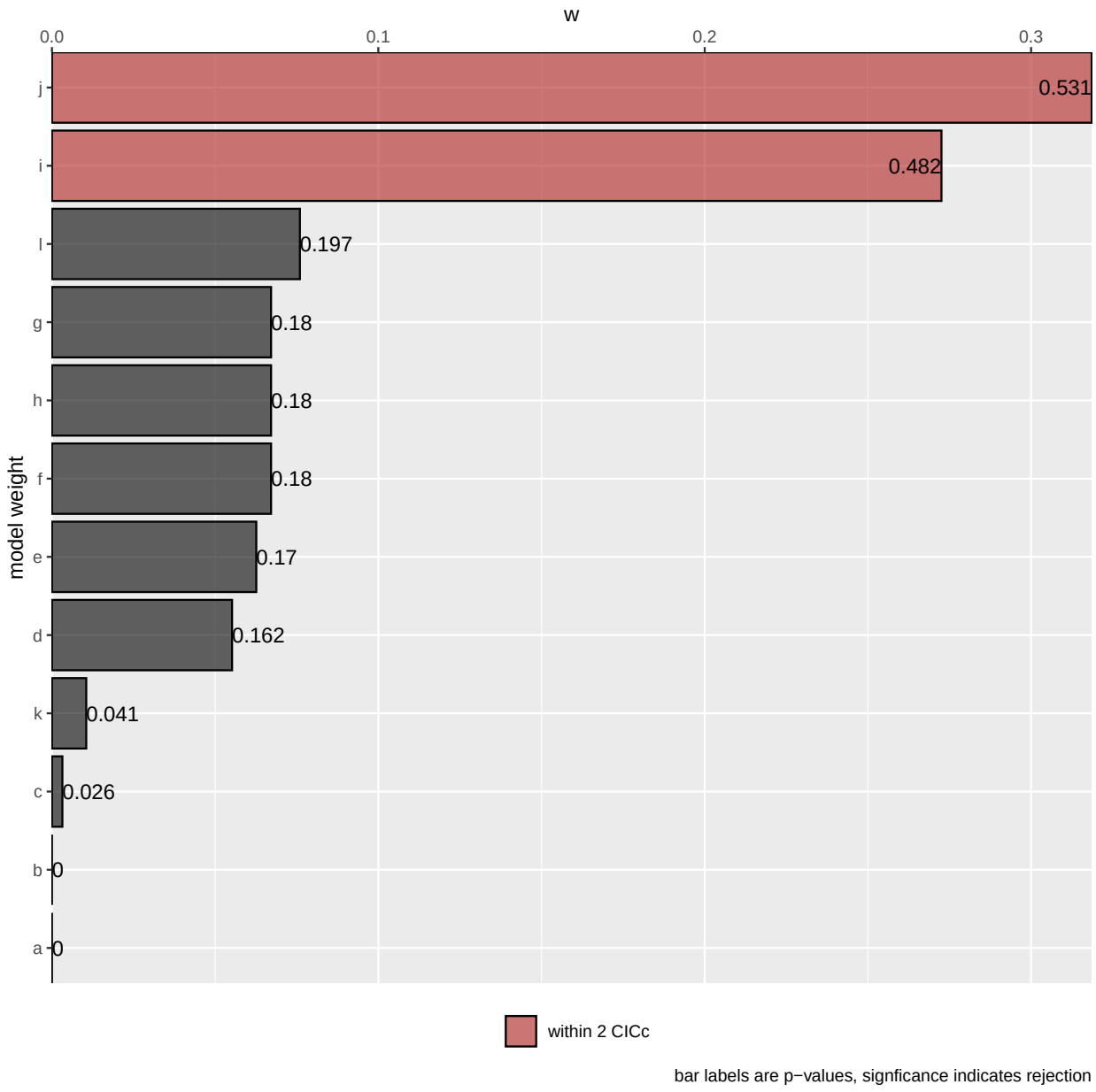


Figure S11: Comparing model support for set of more complex models of causation between four lifestyle traits and pangenome fluidity. Models are ordered by their value of `w`, a measure of model support, and numbers on the bars correspond to overall p-values of the model.

We can see that two models, i and j, have particularly high values of `w`, meaning strong support. Their bars are coloured red because they are within 2CICs of one another, indicating they have a similar level of support. These models are identical except for the direction of causation between species' effect on host and motility.

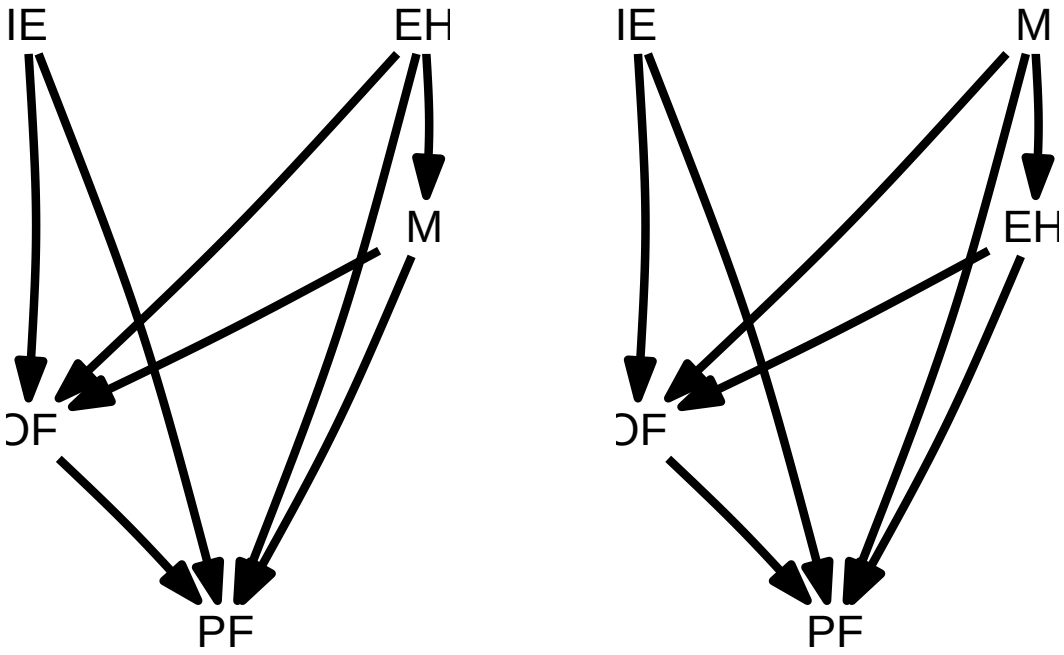


Figure S12: The two models of causation with highest support, model i (first panel) and model j (second panel)

Therefore, the model of causation with the best support is an average of models i and j, which can be viewed below.

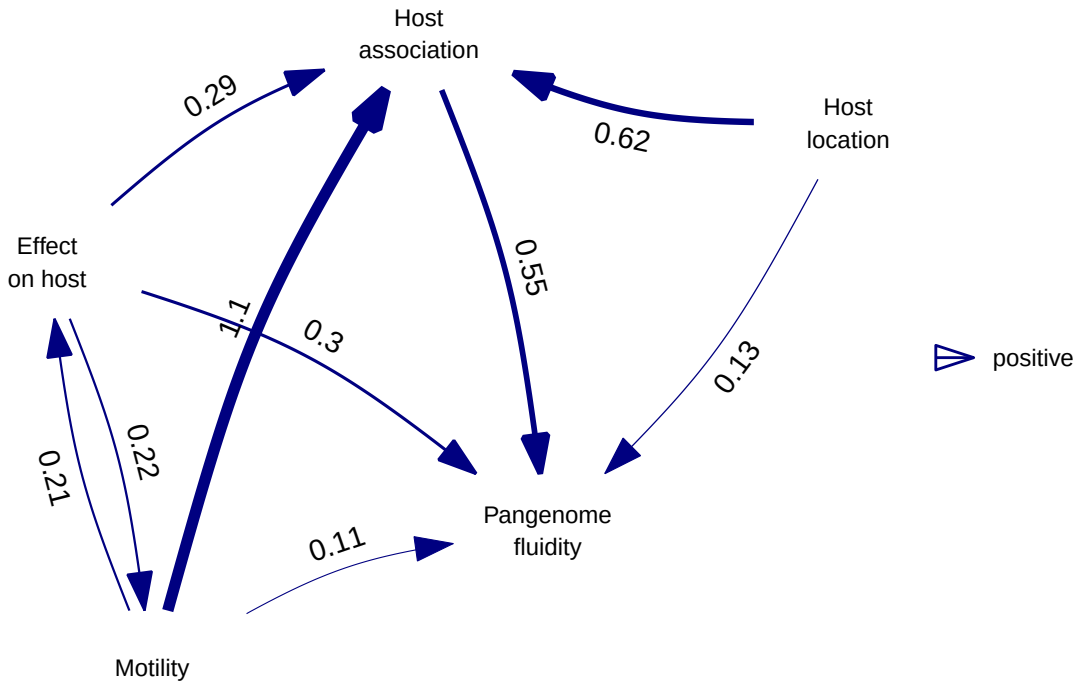


Figure S13: Average best model of causation between four lifestyle traits and pangenome fluidiy

The numbers next to each path correspond to standardized correlation coefficients. This can give an idea of the strength of inferred causation for each path, relative to the other causal links in the model. All the values are positive, meaning that all the paths correspond to an increase in one variable causing an increase in another variable.

We can also view the confidence intervals for each of these coefficients.

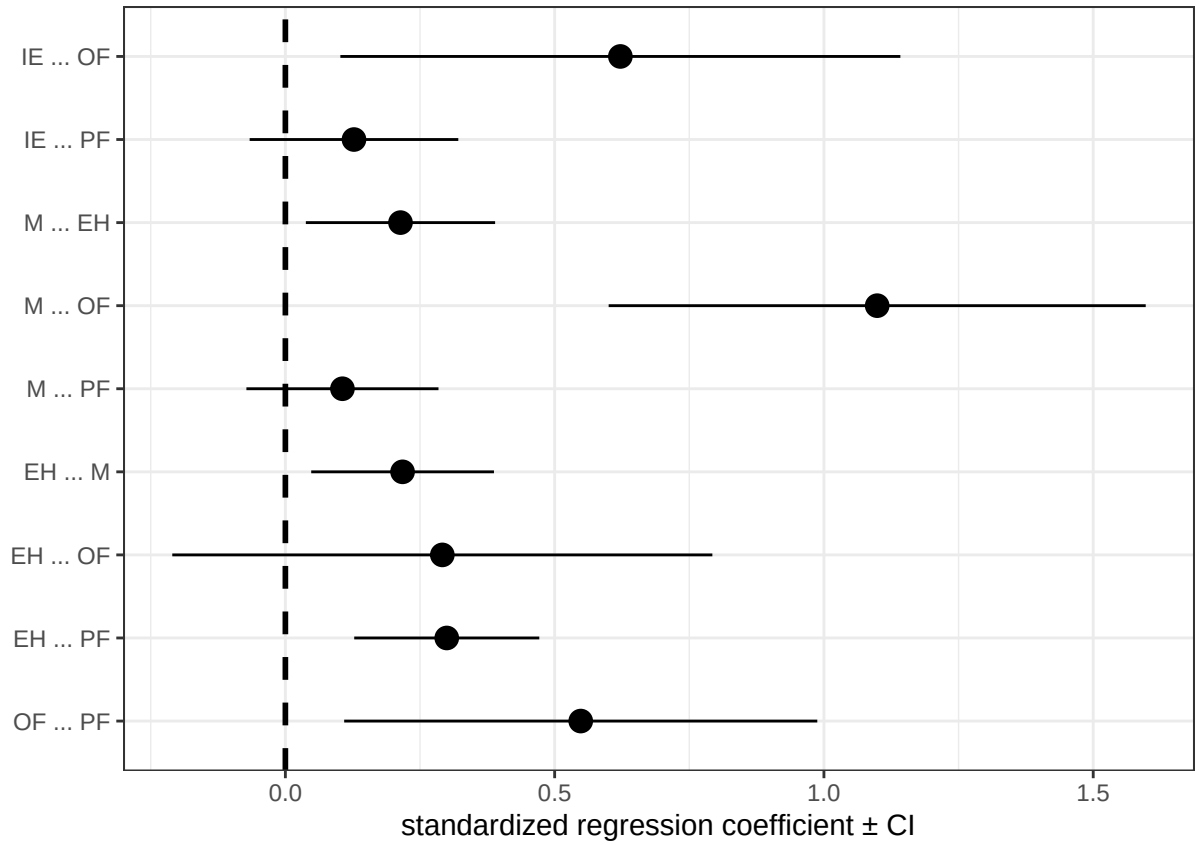


Figure S14: Path coeeficients for the average best model, with 95% confidence intervals.

If the confidence interval does not overlap zero, that means the standardized correlation coefficient is significant to p=0.05 or less. Some of the paths do overlap zero, suggesting weaker support for these paths.

2.7.3 Deletion to the minimal model

We then explored whether the model had better support if one or more of these paths were removed by using a process known as deletion to the minimal model. To do this, we compared a set of models which each had one of the paths removed. We did this for models i and j separately, and also together.

We found that for both models i and j, the original model had the best support. This means that even though some paths have lower support, the model as a whole has best support when all paths are included.

We also compared models i and j, and the models with each of the paths deleted, together within one analysis.

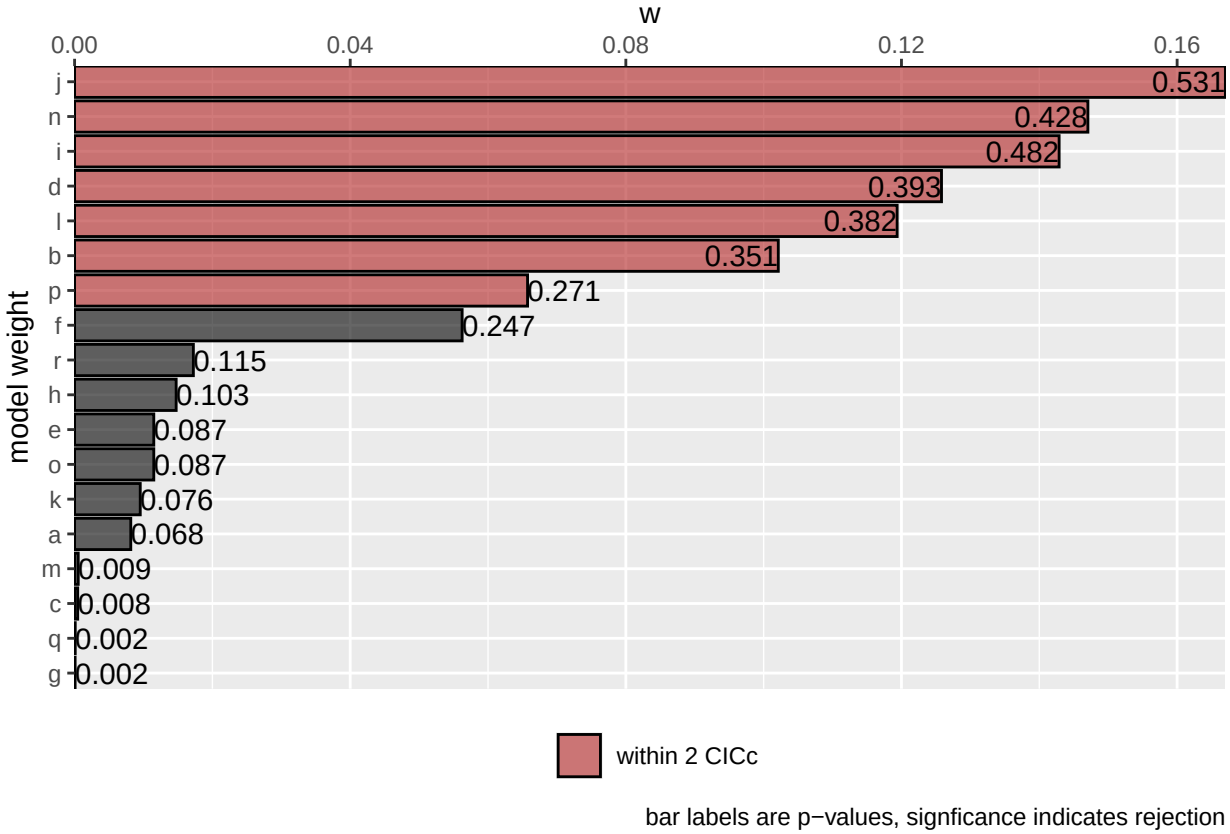


Figure S15: Comparison of support for models i and j, and versions each with one path deleted.

When we considered models i, j, and all versions of i and j with one path deleted, we found there were seven models which were supported to within 2 CICs of one another. Model j had the highest support, with model n second, i in third and d in fourth. Model n is identical to model j, and model d is identical to model i, except that both have the M->PF path removed. This was the path with the lowest standardised coefficient value, suggesting slightly less support for this path than the direct paths to pangenome fluidity from the other lifestyle traits. However, the best model is still one in which all four lifestyle traits have a direct influence on pangenome fluidity.

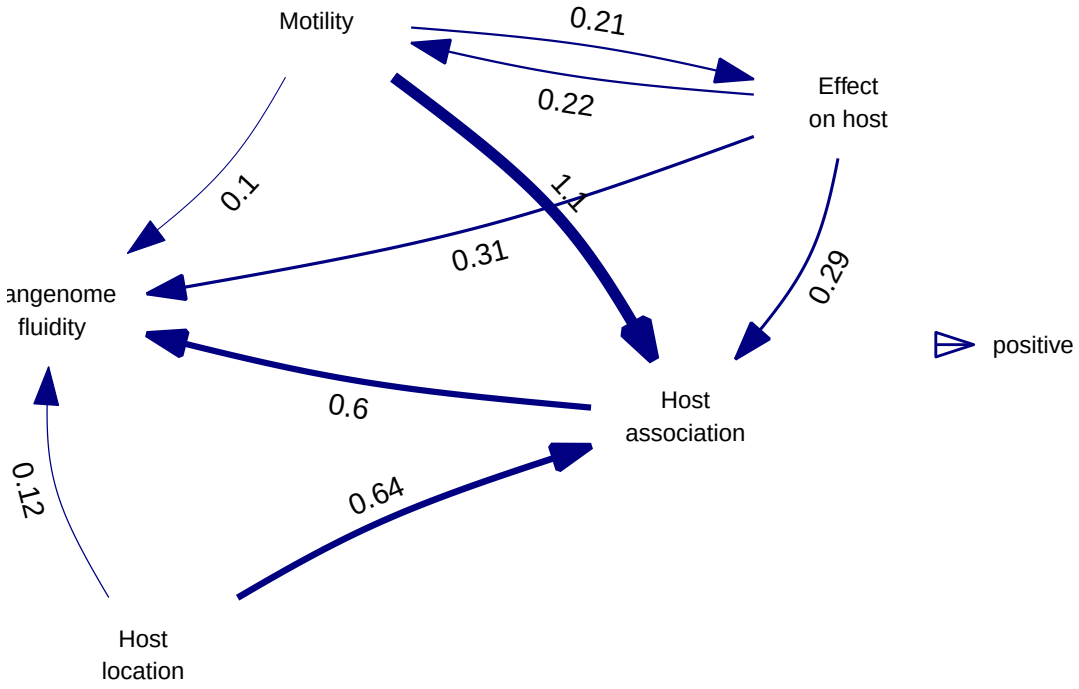


Figure S16: Average best model, combining models i and j, along with 5 models with one oath deleted.

2.8 Phylogentic path analysis with merged intermediate category

In our main path analysis above we included an intermediate catgeory for three of the lifestyle traits. However, for some of our phylogenetic regressions we merged species in these intermediate categories with one of the other two to create binary lifestyle variables. To test the robustness of our results, repeated the above analyses but with binary lifestyle variables.

As in our main analysis, we found that all of the additional simple models, with no paths between lifestyle traits, were rejected (their p-values were all less than 0.05)

As in the previous analysis, we then conducted a path analysis using a more complex set of models, allowing paths between the lifestyle traits.

We found that of these, a single model, model d, had the best support. This model and its correlation coefficients can be viewed in the figures below.

Table S43: Comparison of set of more complex causal models, varying by which lifestyle trait(s) directly influence pangenome fluidity and how they might cause each other.

	w	p	CICc
d	0.3926	0.5870	31.7272
e	0.1167	0.3367	34.1540
g	0.1092	0.3214	34.2866
h	0.1092	0.3214	34.2866
f	0.1092	0.3214	34.2866
c	0.0596	0.2054	35.4988
l	0.0457	0.1696	36.0288
k	0.0251	0.1065	37.2244
i	0.0168	0.0771	38.0311
j	0.0159	0.0737	38.1445
b	0.0000	0.0000	60.5858
a	0.0000	0.0000	76.3654

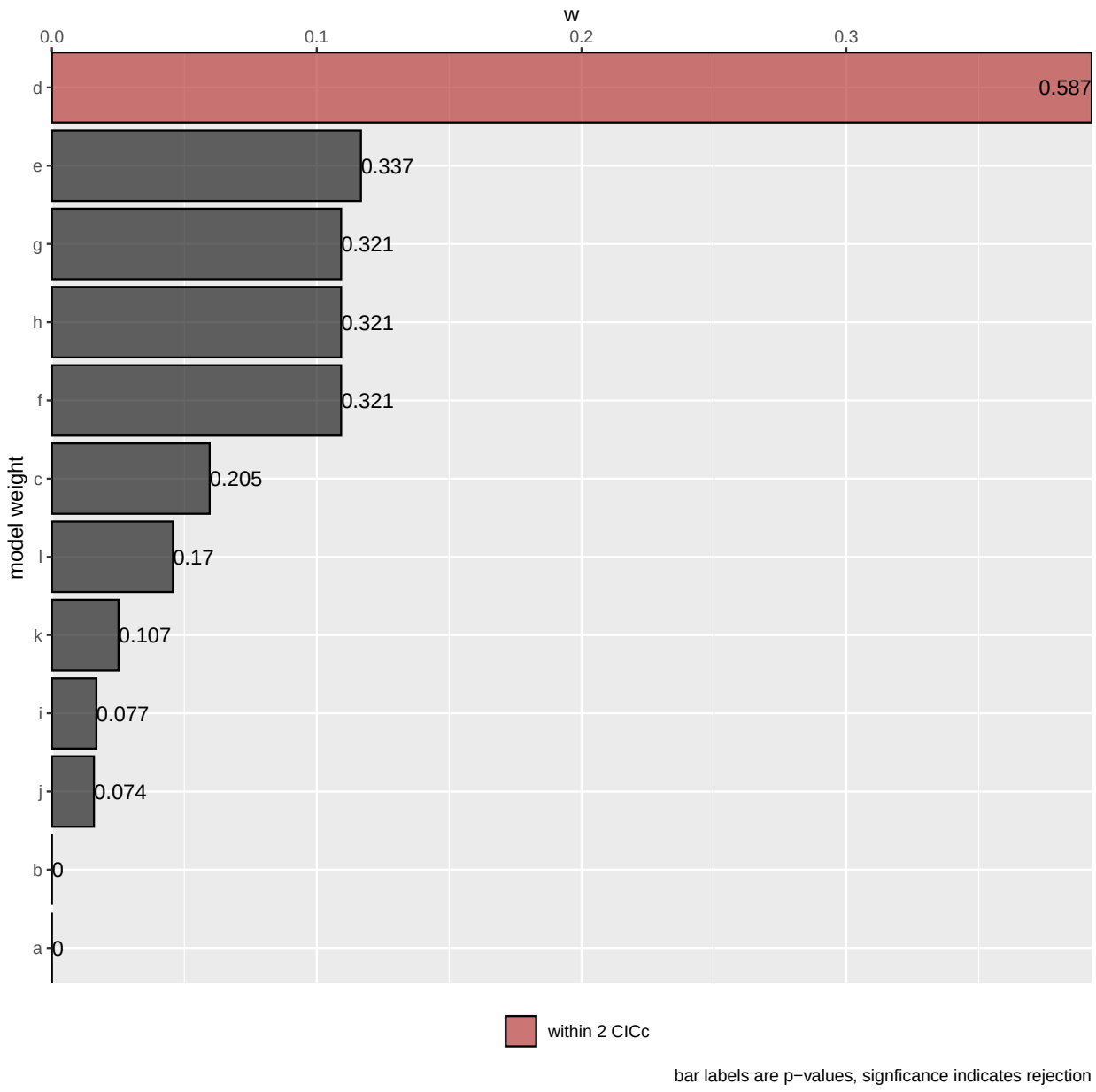


Figure S17: Comparing model support for set of more complex models of causation between four lifestyle traits and pangenome fluidity. Models are ordered by their value of `w`, a measure of model support, and numbers on the bars correspond to overall p-values of the model.

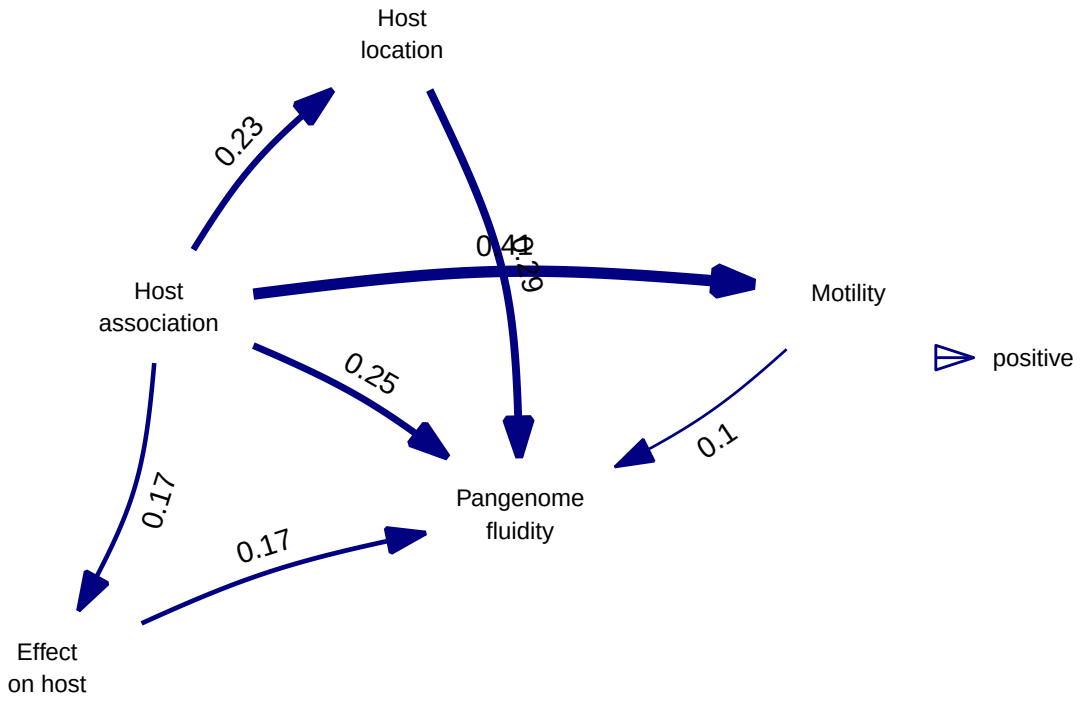


Figure S18: Average best model of causation between four lifestyle traits and pangenome fluidity.

In the average best model from our main analysis, we found evidence that host association was influenced by the other three lifestyle traits.

In this analysis, where we merged the intermediate category to convert those three traits into binary traits, the model with the best support included paths between host association and the other three lifestyle traits, but in the opposite direction as the model in our main analysis.

This provides further evidence that the lifestyle traits are correlated in a manner suggesting that they are coevolving, since a relatively small change of how to code the lifestyle trait can change the direction of influence between pairs of traits.

Finally, we repeated the model deletion step, as in our main analysis.

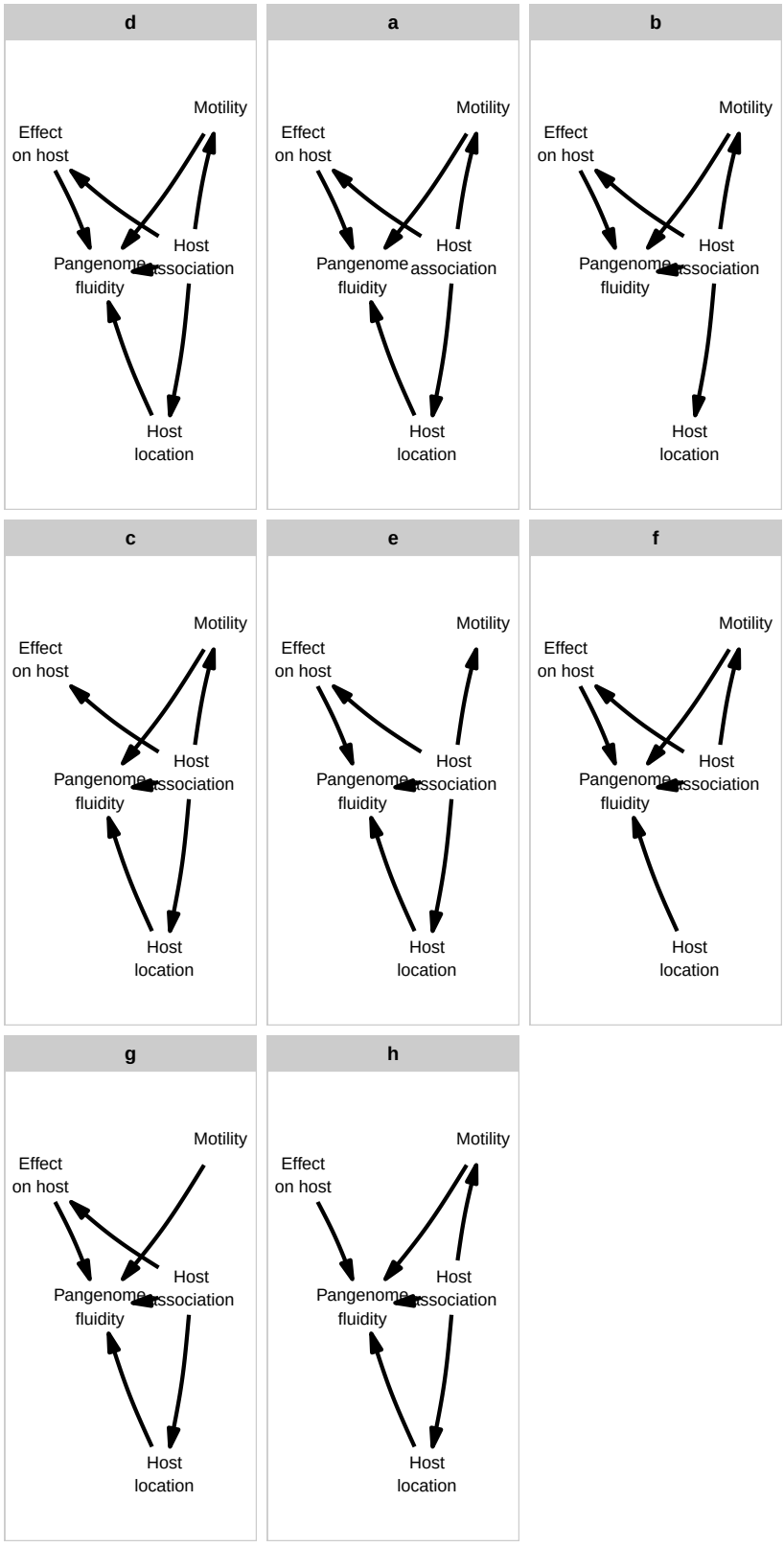


Figure S19: Models each with a path deleted from original (model d).

Table S44: Comparison of support for set of models each with a path deleted from original (model d).

	w	p	CICc
d	0.3807	0.3120	34.1551
e	0.3270	0.2724	34.4590
b	0.2654	0.2437	34.8761
a	0.0212	0.0524	39.9320
f	0.0045	0.0179	43.0409
c	0.0012	0.0067	45.7185
h	0.0001	0.0008	51.1470
g	0.0000	0.0000	58.1807

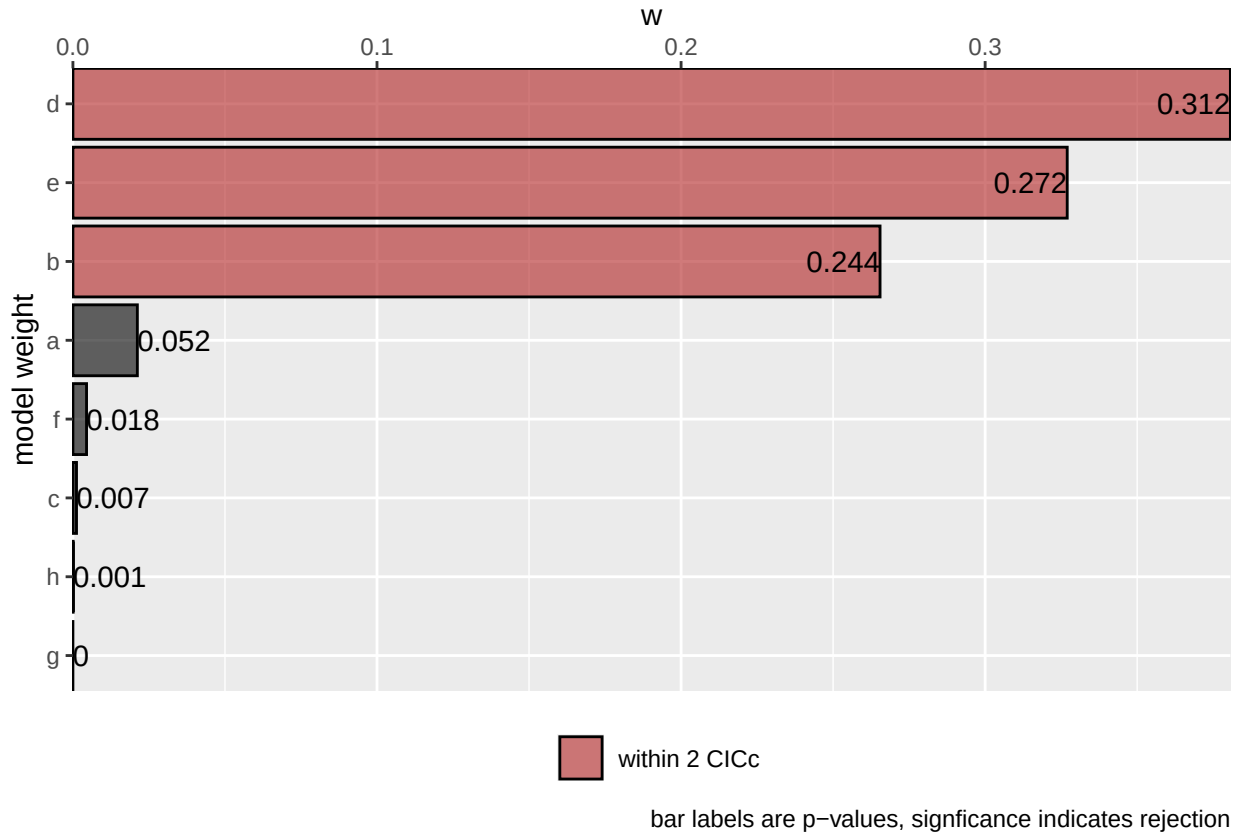


Figure S20: Comparison of support for set of models each with a path deleted from original (model d).

As in our main analysis, we found that the best model was the original model, with no paths deleted. We found that two additional models had support of within 2 CICs of model d. Model e has the path from motility to pangenome fluidity deleted, while model b has the path from host location to pangenome fluidity deleted. This suggests that these paths have weaker support, but that the model is still best when they are included.

We can plot a best model from this analysis where all lifestyle traits are coded as binary, by averaging the three models within 2 CICs of each other, as in the above figure.

This gives the model below.

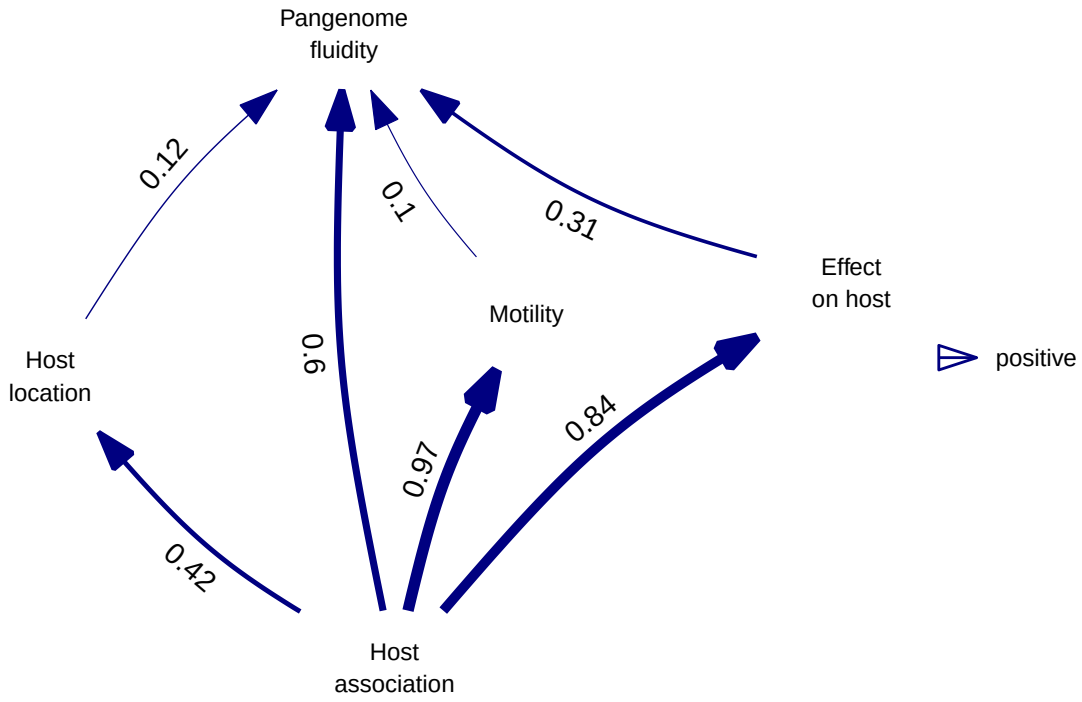


Figure S21: Average of three best models following model deletion, for analysis with all lifestyle traits coded as binary variables.

Considering these results together with the average best model from our main analysis, we find strong evidence that multiple aspects of bacterial lifestyle together influence variation in pangenome fluidity across species. Additionally, we find evidence that the lifestyle traits themselves influence the evolution of each other. The direction of how each influences the other appears dependent on how we code our lifestyle traits (for example, whether we include an intermediate category or treat each trait as binary). However, which pairs of traits were linked by paths between the traits were generally very similar between the two models. This provides further evidence that the traits are co-evolving with one another, likely both influencing the evolution of each other.

For example, the transition from a facultative extracellular host-associated lifestyle to an obligate intracellular lifestyle involves both an increase in frequency of living inside cells, and also a increase in the dependency of the bacteria on its host(s). We find evidence that these likely interact and influence one another as they increase.

2.9 Summary

Overall, we find evidence that each of the four lifestyle traits we examined has a direct influence on a species' pangenome fluidity. Specifically, we find that species which are facultatively host associated, live outside host cells, are mutualists and are motile have a higher pangenome fluidity compared to species which are obligately host-asociated, live inside host cells, are pathogens and are non-motile.

These correlations are likely because each of these lifestyles influences the variety of genes available for individuals of a species to acquire, and also the variety of environments individuals of a species will encounter and need different sets of genes to live in. Therefore, multiple bacterial lifestyle traits influence pangenome fluidity because they influence rates of gene gain and loss across individual bacteria of a species.

3 What about other factors in addition to lifestyle: genome size and effective population size?

Two additional factors correlate with pangenome fluidity, which are not lifestyle traits but instead genome characteristics: genome size and effective population size. Species with larger genome sizes and larger effective population sizes are predicted to have higher pangenome fluidity.

Genome size could influence genome fluidity by limiting the number of genes which can vary across individuals of the same species: each species has a set of core genes which all individuals need to survive, so if the genome size is not much bigger than this number, the average proportion of genes which differs between individuals will be low, meaning pangenome fluidity will be low.

Effective population size, which measures the number of individuals in the current generation which will leave descendants in a distant generation, is important for how genes spread across generations. Genetic drift has a larger influence in species with low effective population sizes, reducing the chances that genes with small but beneficial genes will spread, while also slowing down the rate that genes which are slightly deleterious, such as those no longer required, will be purged from the population. Together, this could influence genome fluidity by limiting any local adaptation of subpopulations of a species in a particular environment, meaning pNgenome fluidity is reduced.

3.1 Genome size and effective population size correlate with pangenome fluidity

First, we examined whether both variables were correlated with pangenome fluidity across our species.

We calculated the mean number of genes in genomes of all 126 species in our dataset as a measure of genome size. We found that, as predicted and as has been observed previously, that species with larger genomes had higher pangenome fluidity.

Table S45: Results from a MCMCglmm with pangenome fluidity as the response variable, genome size as the fixed effect, and phylogeny as a random effect.

	post.mean	l-95% CI	u-95% CI	eff.samp	pMCMC	signif.
(Intercept)	0.16596059	0.086128010	0.24532614	4700	0.0012766	**
genome_size	0.00001805	0.000007827	0.00002877	4700	0.0008511	***

	R-squared value
Fixed effect	0.06444
Random effect	0.53484
Total model	0.59928

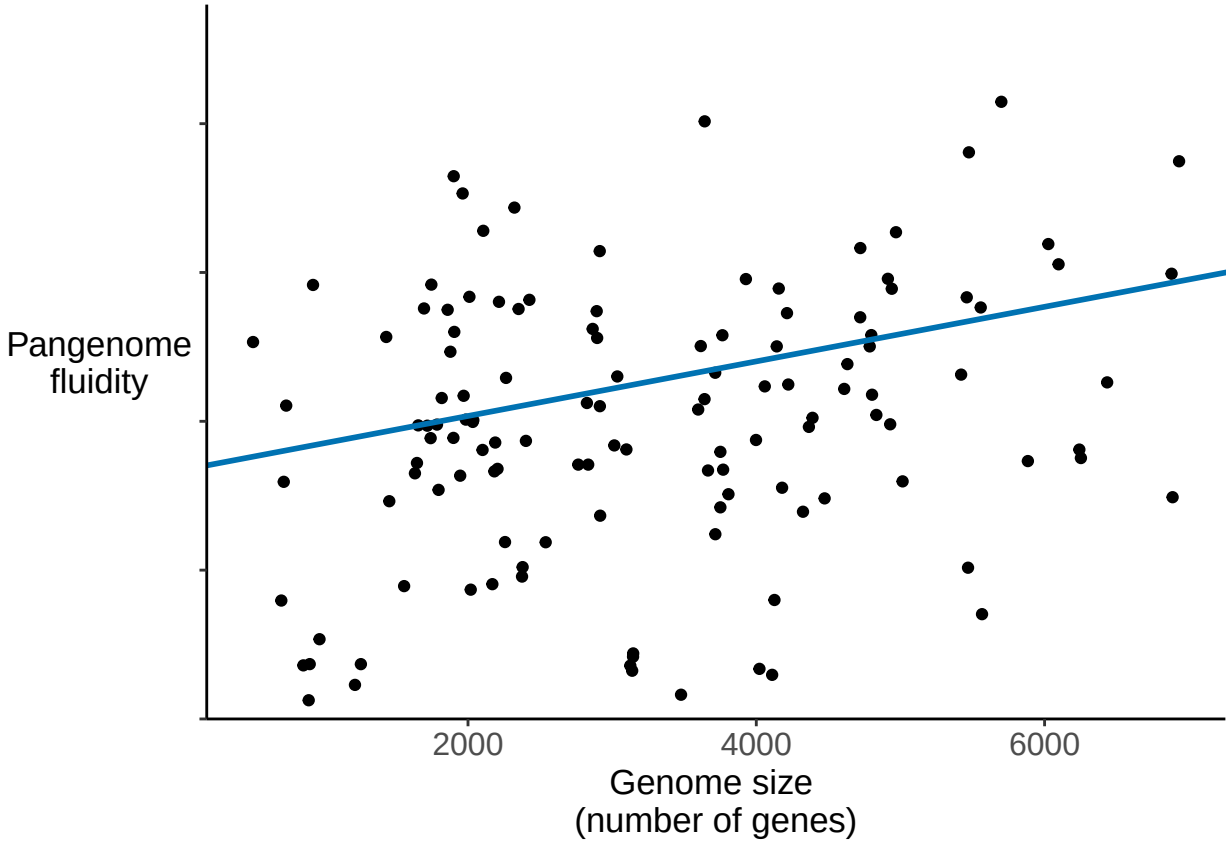


Figure S22: Plot showing correlation between a species' average number of genes in their genomes and their pangenome fluidity. The line is the slope and intercept from the MCMCglmm analysis in the above table. N=126

We used previously calculated estimates of effective population size, available for 77 of our species, which were based on dN/dS ratios in a set of universal genes (L.M. Bobay, H. Ochman, Factors driving effective population size and pan-genome evolution in bacteria. BMC Evolutionary Biology. 18, 153 (2018)). We found that, as predicted and as has been found previously, species with larger effective population sizes had higher pangenome fluidity.

Table S46: Results from a MCMCglmm with pangenome fluidity as the response variable, effective population size as the fixed effect, and phylogeny as a random effect.

	post.mean	l-95% CI	u-95% CI	eff.samp	pMCMC	signif.
(Intercept)	0.1030692	0.0170474	0.1999065	3659	0.0259574	*
Ne_small	0.0003523	0.0001432	0.0005514	4100	0.0002128	***

	R-squared value
Fixed effect	0.08568
Random effect	0.46281
Total model	0.54849

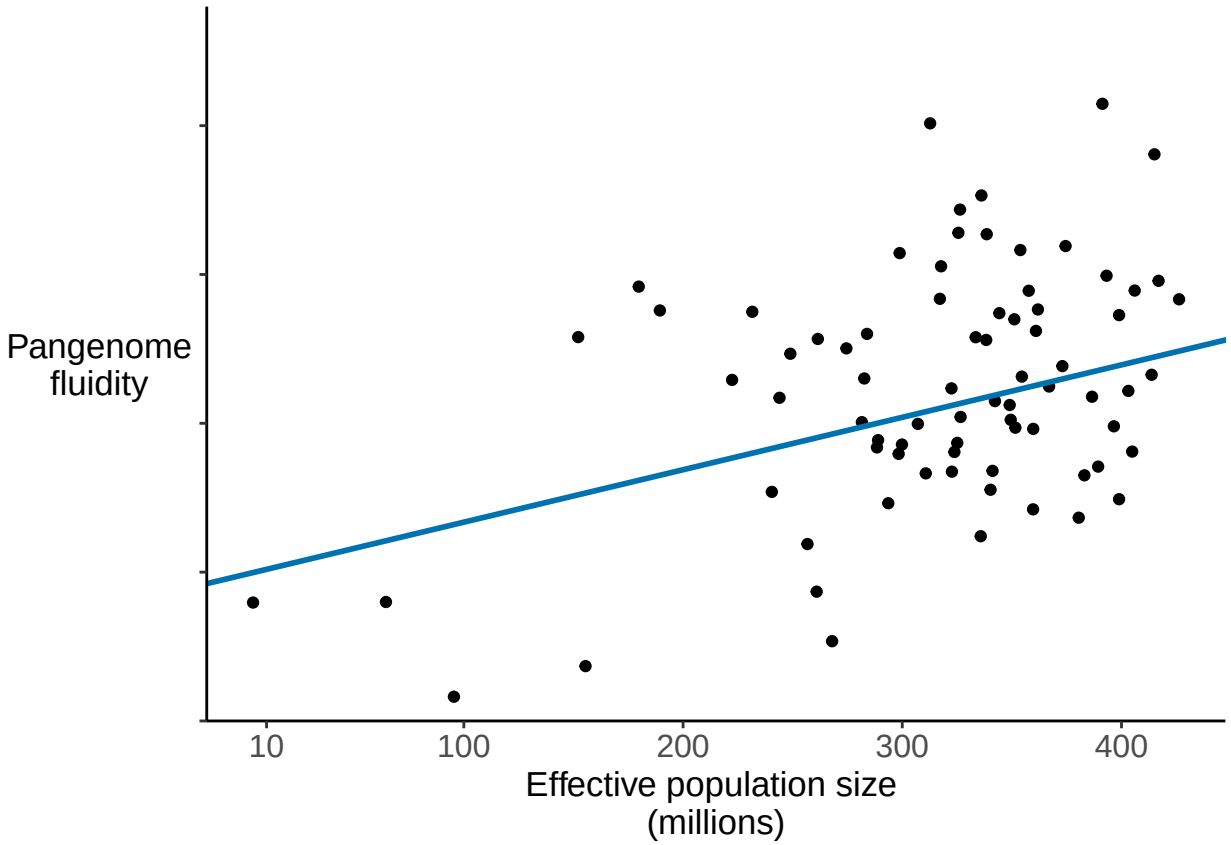


Figure S23: Plot showing correlation between a species' effective population size and their pangenome fluidity. The line is the slope and intercept of the MCMCglmm analysis in the above table. N=77.

3.2 Phylogenetic path analysis

It is unclear whether the correlation between both factors and pangenome fluidity is because each is directly influencing pangenome fluidity, or because some other factor is influencing one or both. Lifestyle could be important, potentially influencing both factors, and so may also explain why such factors correlate with pangenome fluidity.

To examine this, we used phylogenetic path analysis to investigate which of lifestyle, genome size and effective population size are most important for causing changes in genome fluidity. We did this for the 75 species for which we had data on effective population sizes and also all four lifestyle traits.

To characterise species' lifestyle, we combined our four within-host lifestyle traits into one variable. In our main analyses, we did this by coding the category with the lowest lifestyle variability as 0 and the category with highest as 1. For those with 'both' categories, we coded this as 0.5. We then summed these four values for each species, giving a single measure of lifestyle variability. The variable had a minimum value of 0, which would correspond to an obligate, intracellular, pathogenic, non-motile species, predicted to have the least variable lifestyle. Conversely, the maximum value of the variable is 4, corresponding to a facultative, extracellular, mutualistic, motile species, which would have the most variable lifestyle.

This single lifestyle variable was significantly correlated with pangenome fluidity.

Table S47: Results from a MCMCglmm with pangenome fluidity as the response variable, lifestyle vairability as the fixed effect, and phylogeny as a random effect.

	post.mean	l-95% CI	u-95% CI	eff.samp	pMCMC	signif.
(Intercept)	0.10563	0.02423	0.18093	4700	0.0097872	**
LS	0.04605	0.02633	0.06451	4700	0.0002128	***

	R-squared value
Fixed effect	0.1975
Random effect	0.4001
Total model	0.5975

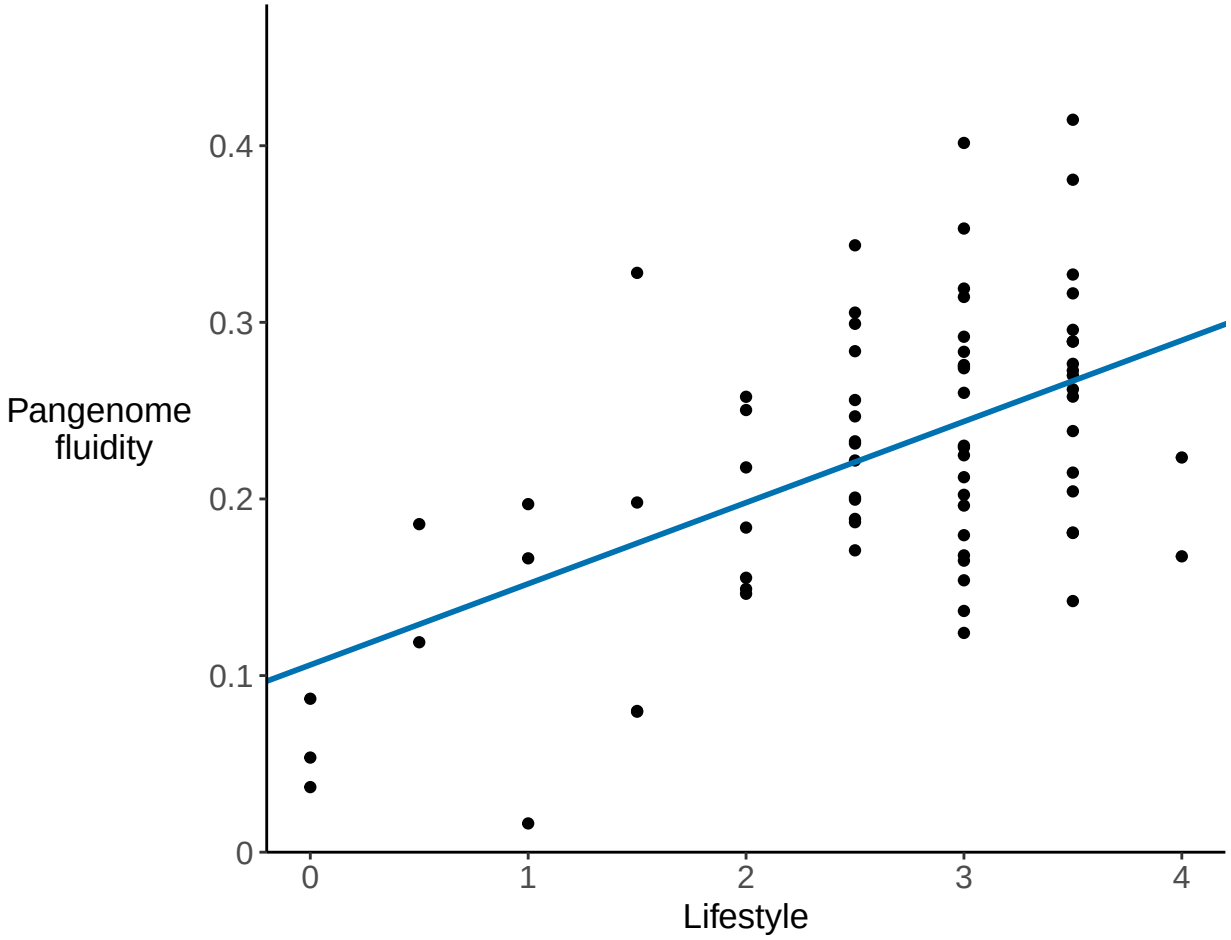


Figure S24: Plot showing correlation between lifestyle as a single variable and species' pangenome fluidity. The line is the slope and intercept from the MCMCglmm analysis in the table above.

3.2.1 Simple models with no paths between the three factors

We first compared a set of simple models, with no links between the three potential causal factors Instead, we varied only whether each of the three factors had a direct influence on pangenome fluidity.

We again used `phylo_path()` function to compare support for this set of simple models. In this analysis we

used the Brownian Motion model of evolution, since the models had the highest support with this model of evolution, and the default method for the evolution of binary traits (which applies to host assocation only), `"logistic_MPLE"` (maximizes the penalized likelihood of the logistic regression).

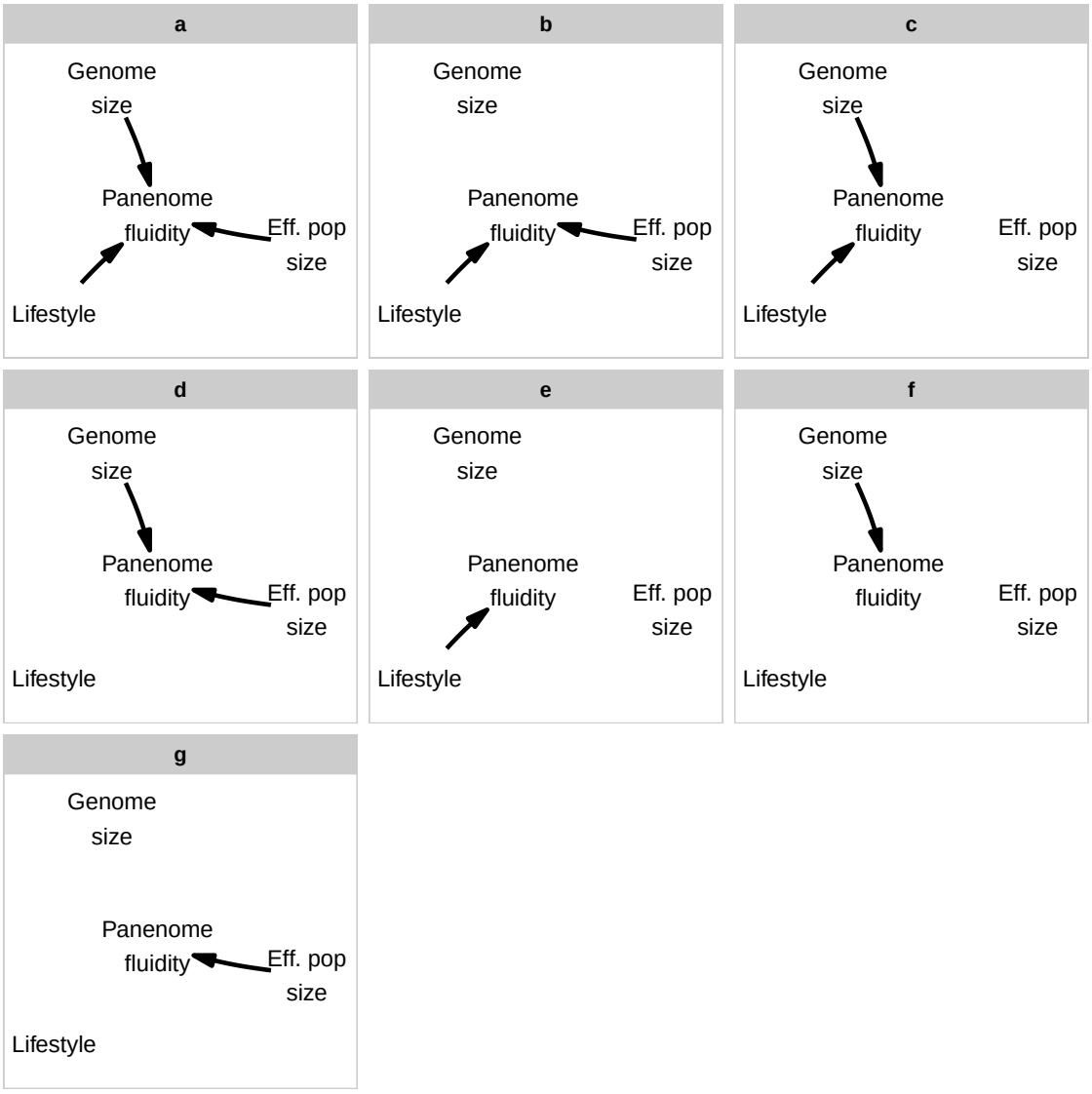


Figure S25: Set of simple models, varying by which of lifestyle, effective population size and genome size causes pangenome fluidity.

We found that a single model had the highest support. This model (model c) includes a direct causal influence of lifestyle and genome size on pangenome fluidity, but not effective population size. Three models were rejected as possible models of causation, meaning they had a p-value of less than 0.05. These were the three models which did not include a direct influence of lifestyle on pangenome fluidity.

Table S48: Details of of support for a simple set of models varying by which of lifestyle, effective population size and genome size causes pangenome fluidity.

	w	p	CICc
c	0.7136	0.5364	20.2376
a	0.2261	0.3336	22.5359
e	0.0429	0.1324	25.8597
b	0.0169	0.0699	27.7240
f	0.0003	0.0059	35.5706
d	0.0001	0.0019	37.7201
g	0.0000	0.0001	46.9693

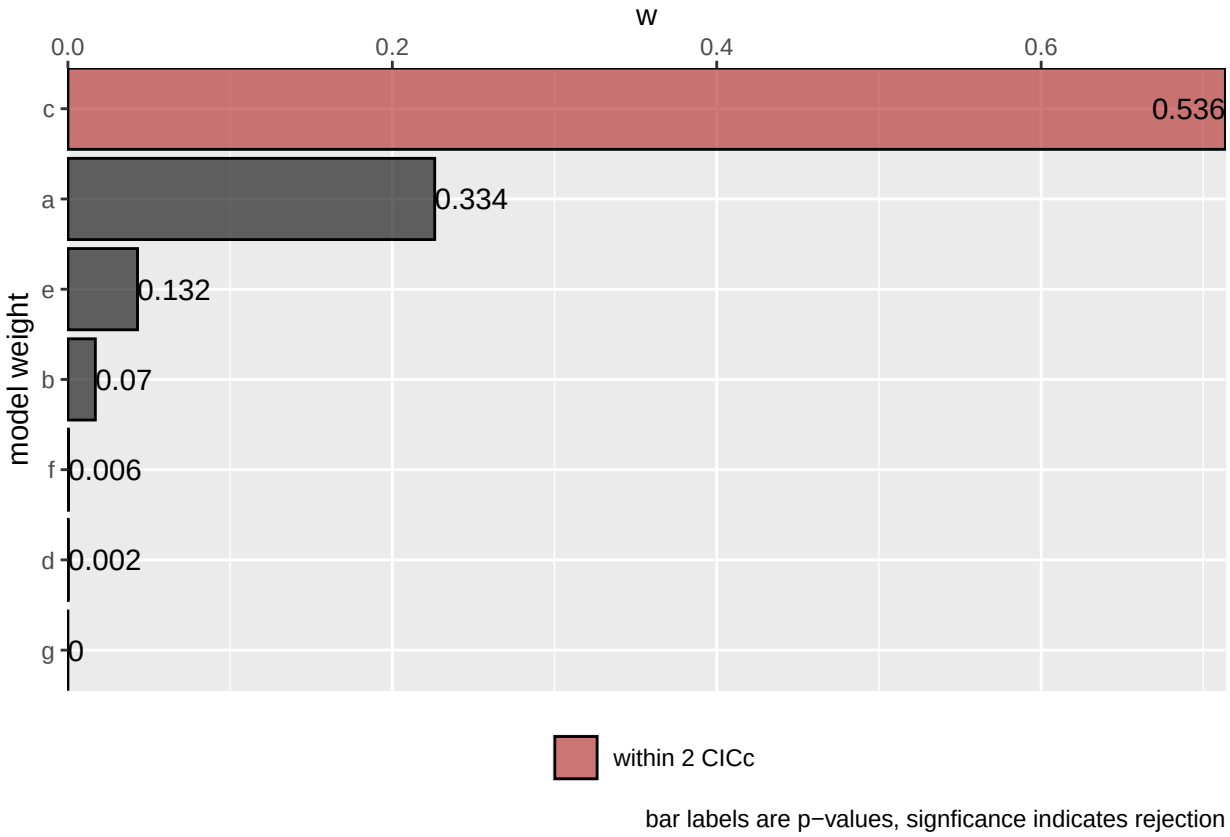


Figure S26: Comparison of support for a simple set of models varying by which of lifestyle, effective population size and genome size causes pangenome fluidity.

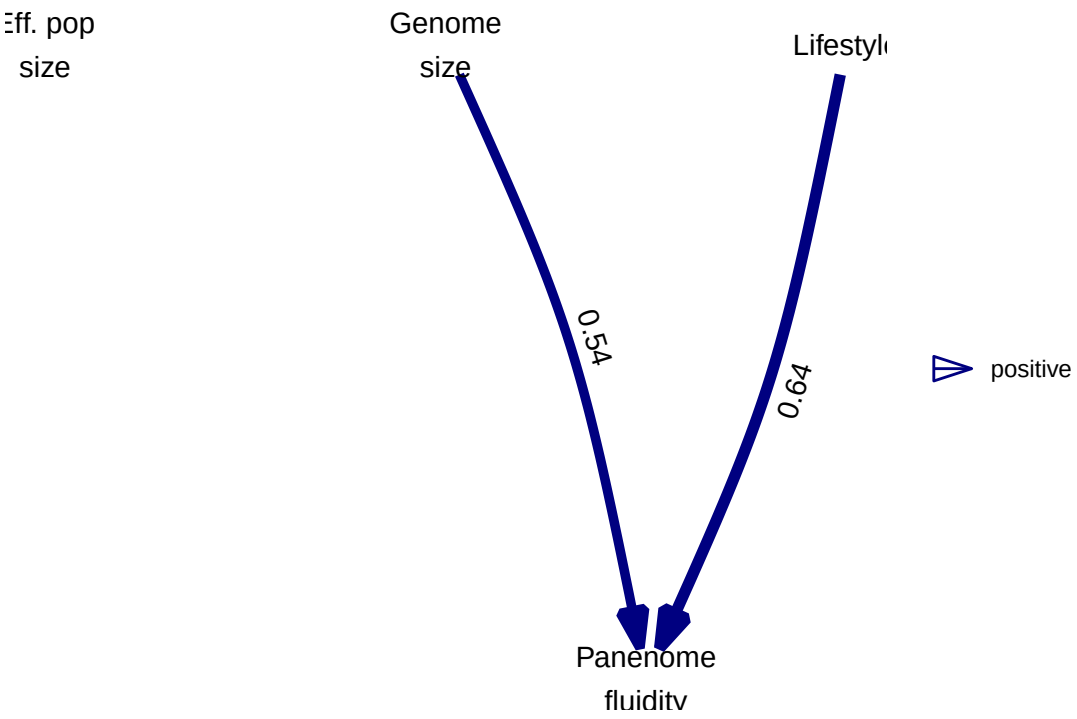


Figure S27: Model with best support, out of a set of simple models.

However, effective population size still clearly correlates with genome fluidity, suggesting there is some unresolved causal structure. This could be due to indirect links between the three traits, for example by lifestyle also causing effective population size variation.

3.2.2 Complex models with paths between the three factors

To explore this, we compared a more complex set of models, allowing for potential links between the three factors. Paths we included and tested were: Lifestyle->Eff pop size, Lifestyle-> genome size, and Eff pop size->genome size; each of these are biologically plausible and could explain why effective population size is correlated with both factors, in addition to genome fluidity.

We compared this set of models, varying the inclusion of each of these paths in addition to the simple models in Figure S25.

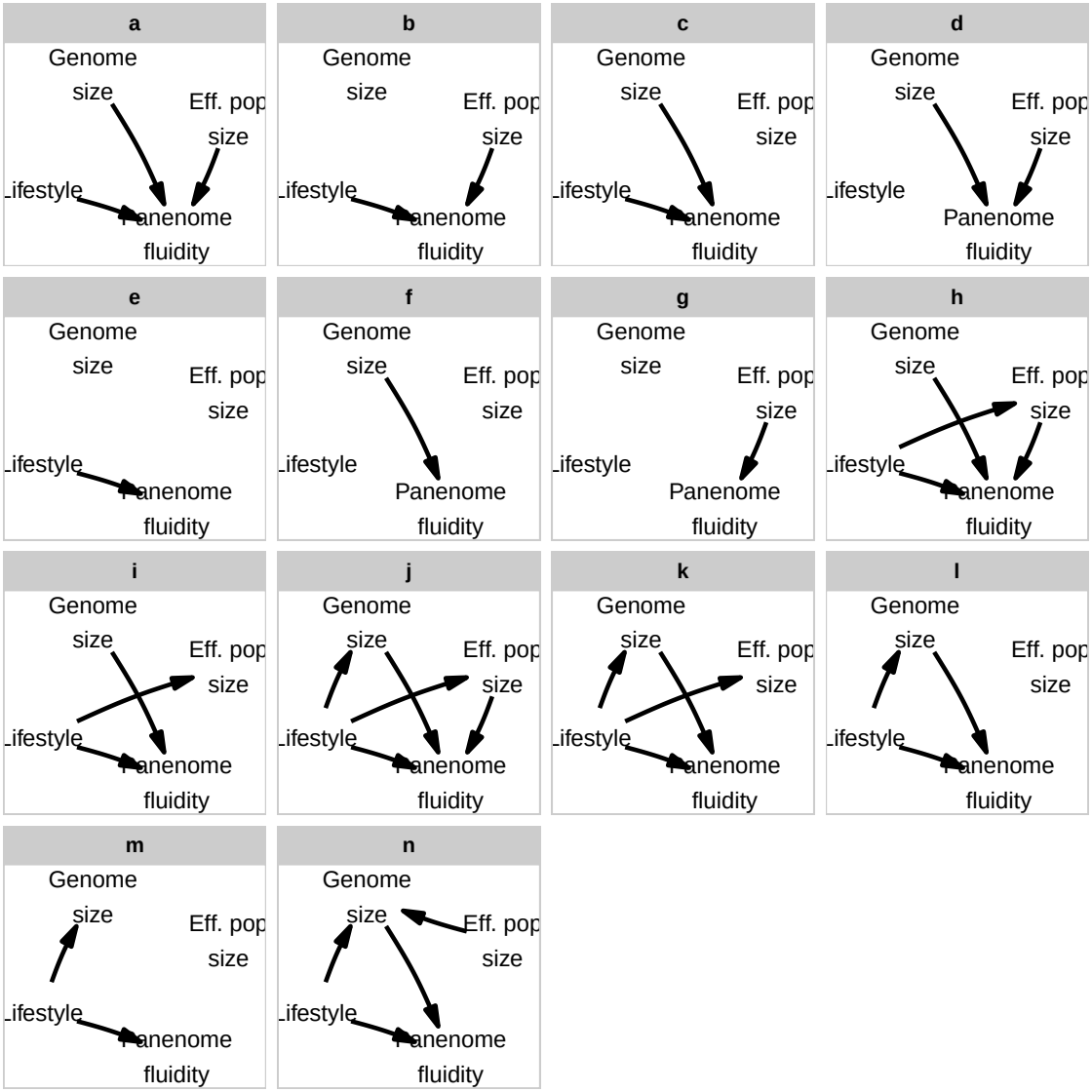
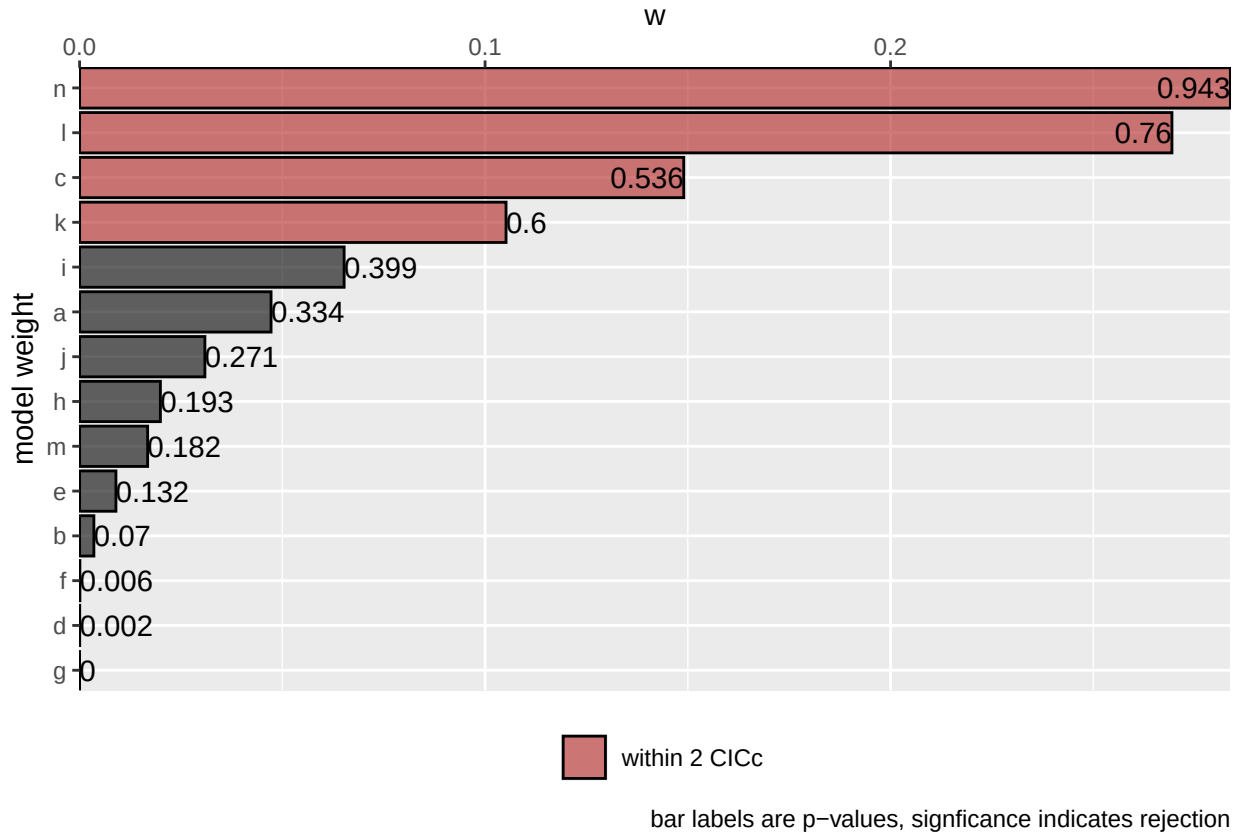


Figure S28: Set of simple and more complex models, varying by which of lifestyle, effective population size and genome size causes pangenome fluidity, and also how those factors might influence each other.

Of these models, four had high and similar support (within 2CICs of each other). The best was model n, which was similar to model c but also with NE->GS and LS->GS. The second best, model l, is identical except it has no NE->GS.

Table S49: Comparison of set of more complex causal models, varying by which of lifestyle, genome size and effective population size directly influence pangenome fluidity, and how they might cause each other.

	w	p	CICc
n	0.2838	0.9428	18.9489
l	0.2694	0.7597	19.0531
c	0.1490	0.5364	20.2376
k	0.1052	0.6001	20.9343
i	0.0652	0.3993	21.8893
a	0.0472	0.3336	22.5359
j	0.0309	0.2706	23.3837
h	0.0199	0.1933	24.2614
m	0.0168	0.1816	24.6046
e	0.0090	0.1324	25.8597
b	0.0035	0.0699	27.7240
f	0.0001	0.0059	35.5706
d	0.0000	0.0019	37.7201
g	0.0000	0.0001	46.9693



We then averaged these four models together. This gives the overall best model, shown in the below figure. This was the basis for Figure 4 in the main text.

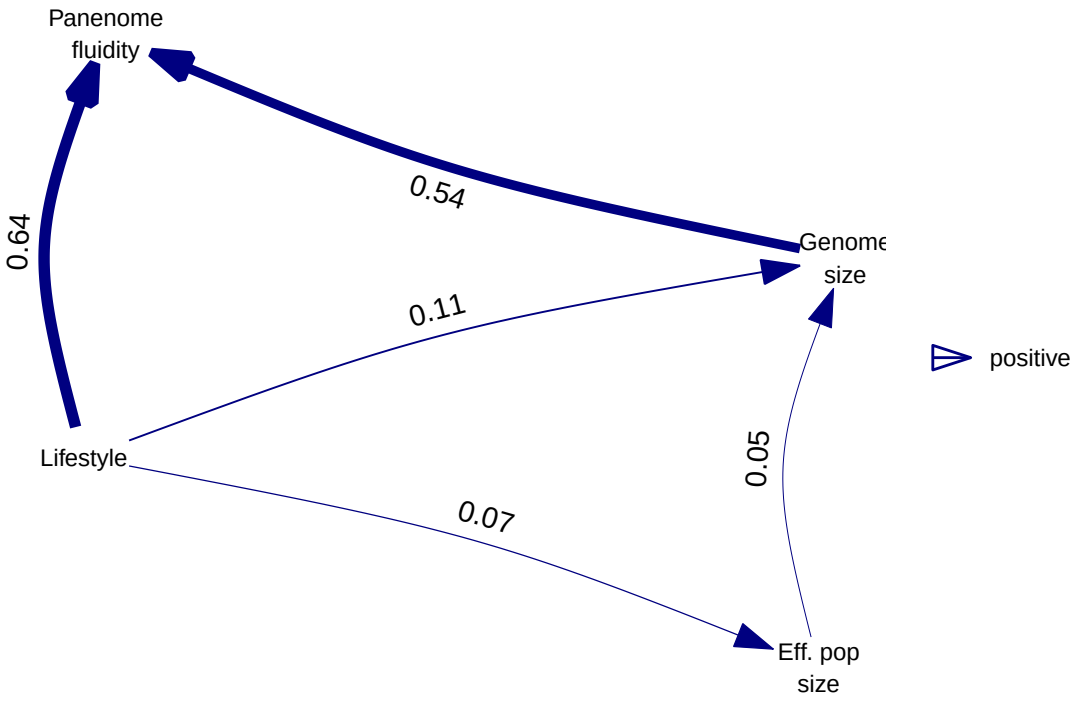
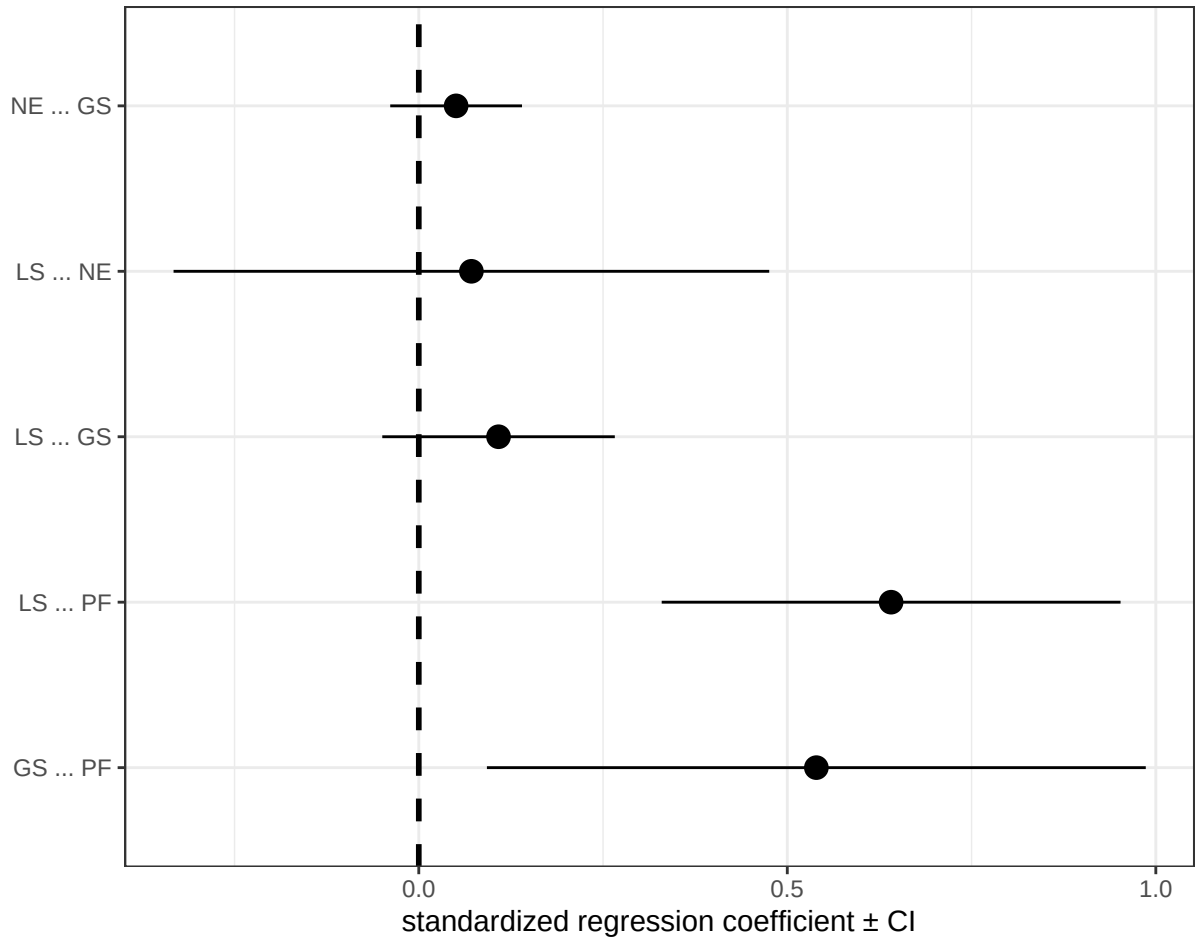


Figure S29: Best model of causal relationships between lifestyle, genome size, effective population size and pangenome fluidity; an average of four models with similar structure. This is the same model as Figure 4 in the main text.

We can also view the 95% confidence intervals for each of the path coefficients in this average best model.



We can also examine this result by setting a path's standardised correlation coefficient to zero if it is not in present in one of the four models with similar support. This weighs the value of the coefficient by how often it is present. This does not affect the paths from genome size and lifestyle to pangenome fluidity, since they are present in all four models. However, it does reduce the size of the correlation coefficients for the remaining paths, particularly those factors influencing and that are influenced by effective population size. This provides further evidence that the influence of lifestyle and genome size on pangenome fluidity are the most important paths in the model.

We present the result without these weightings in the main text, in order to report the original correlation coefficient values, but include this here for further detail.

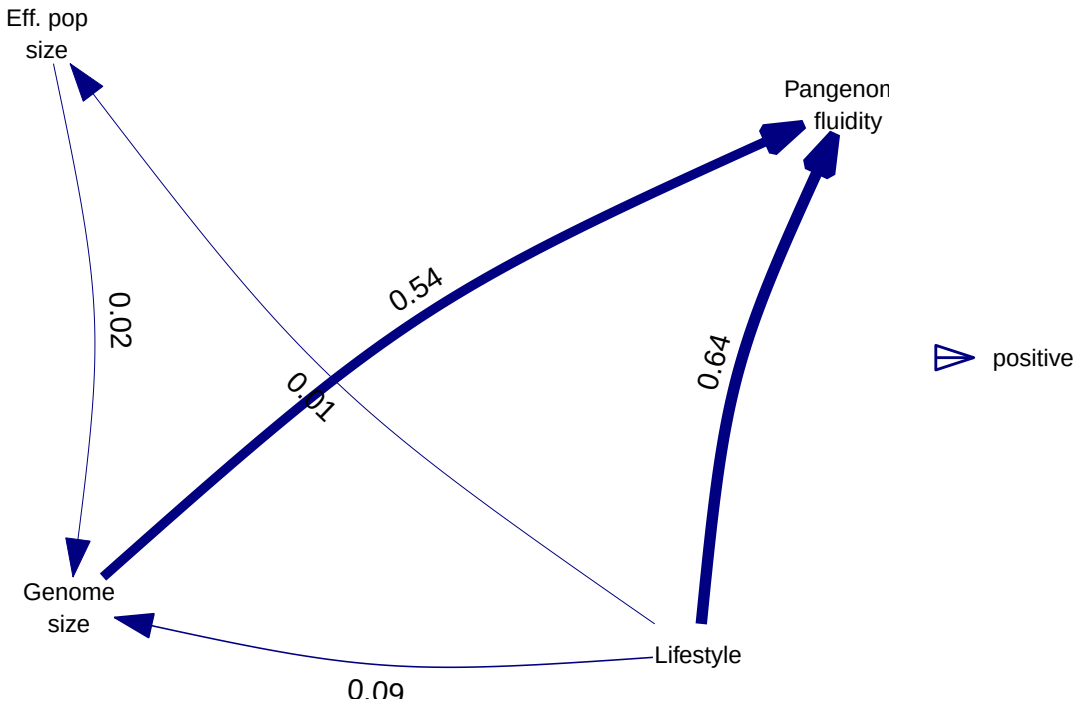


Figure S30: Best model of causal relationships between lifestyle, genome size, effective population size and pangenome fluidity; an average of four models with similar structure, with correlation coefficients weighted by setting absent paths to zero when averaging models.

3.3 Other measures of lifestyle

Next, we explored whether these results were robust to other features and/or measures of lifestyle. We did this in several ways.

3.3.1 Single lifestyle trait but with merged intermediate category

First, we did the same analysis but coded the single lifestyle trait slightly differently. For three of our lifestyle traits we had an intermediate category we coded as 0.5 for that trait. Here, we combined species in this intermediate category with species in one of the other two categeories to create a binary variable, as we did for some of the phylogenetic regression analyses.

This alternative single lifestyle variable was also significantly correlated with pangenome fluidity.

Table S50: Results from a MCMCglmm with pangenome fluidity as the response variable, lifestyle vairability as the fixed effect (calculated with no intermediate/both category), and phylogeny as a random effect.

	post.mean	l-95% CI	u-95% CI	eff.samp	pMCMC	signif.
(Intercept)	0.1251	0.04126	0.19628	4461	0.0025532	**
LS	0.0426	0.02451	0.06213	4700	0.0002128	***

	R-squared value
Fixed effect	0.1764
Random effect	0.3965
Total model	0.5730

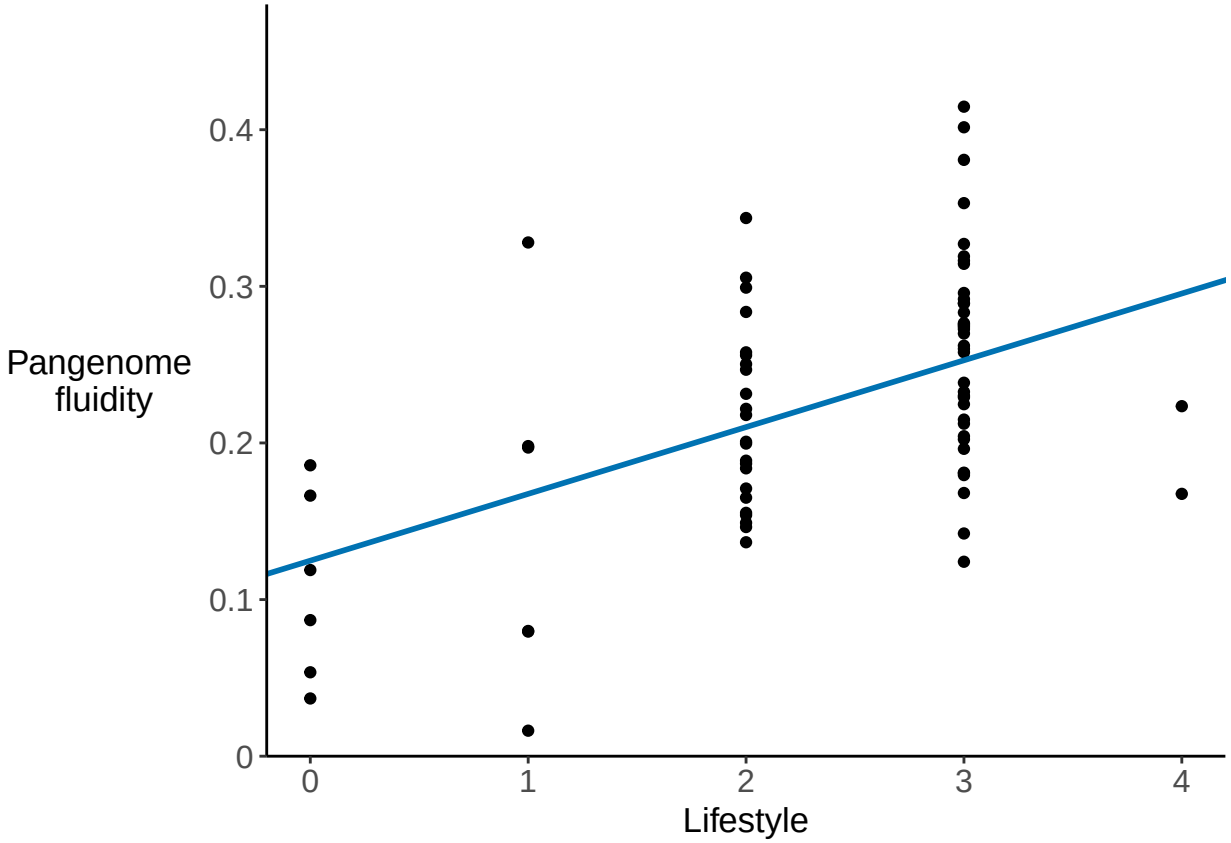


Figure S31: Plot showing correlation between lifestyle as a single variable (calculated with no intermediate/both category) and species' pangenome fluidity. The line is the slope and intercept from the MCMCglmm analysis in the table above.

Table S51: Details of of support for a simple set of models varying by which of lifestyle, effective population size and genome size causes pangenome fluidity.

	w	p	CICc
c	0.6008	0.2335	23.7069
a	0.3557	0.1689	24.7551
e	0.0304	0.0428	29.6755
b	0.0111	0.0181	31.6826
f	0.0015	0.0057	35.6928
d	0.0004	0.0015	38.3504
g	0.0000	0.0001	45.9330

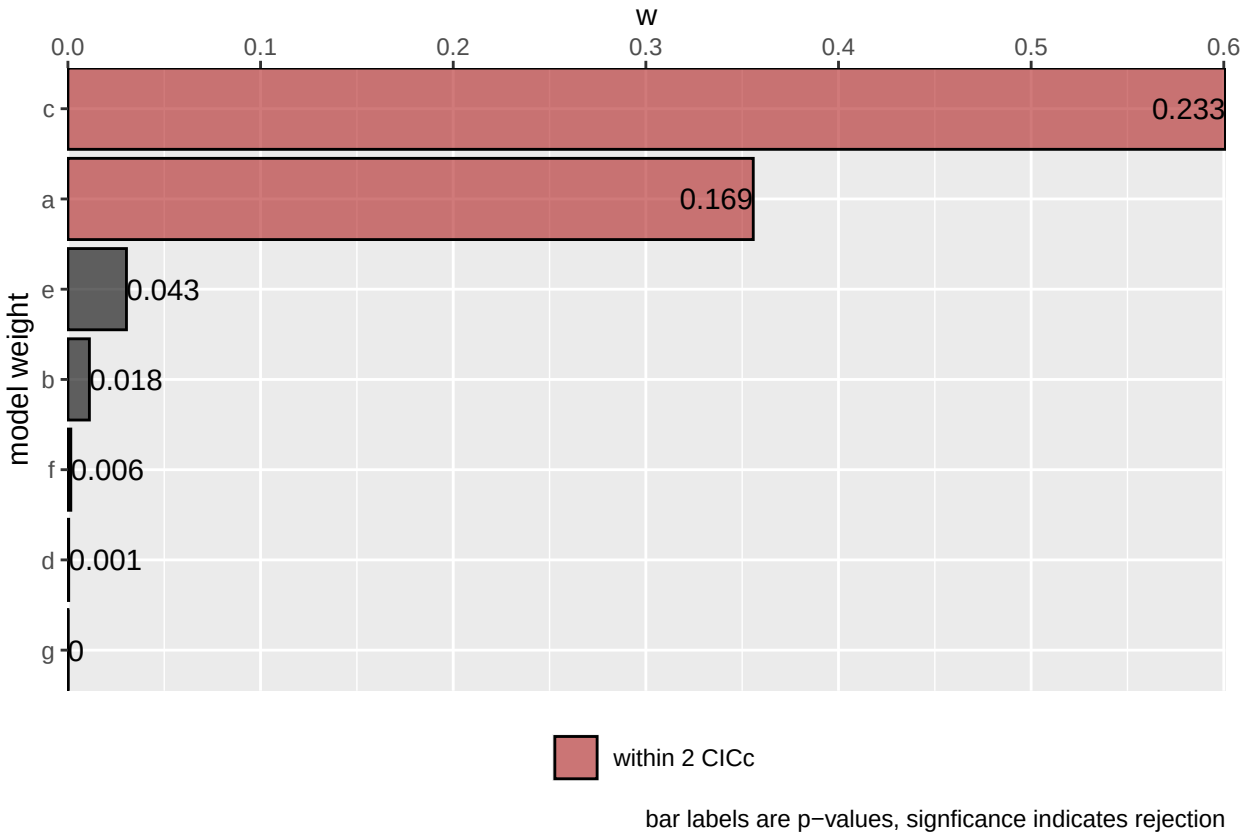


Figure S32: Comparison of support for a simple set of models varying by which of lifestyle (calculated with no intermediate/both category), effective population size and genome size causes pangenome fluidity.

We again found that model c had the best support, with support with model a being within 2CICcs, indicating similar support. We averaged these two simple models together to give the below result. This gives similar support for a direct causal role of lifestyle and genome size, and less support for a direct role of effective population size.

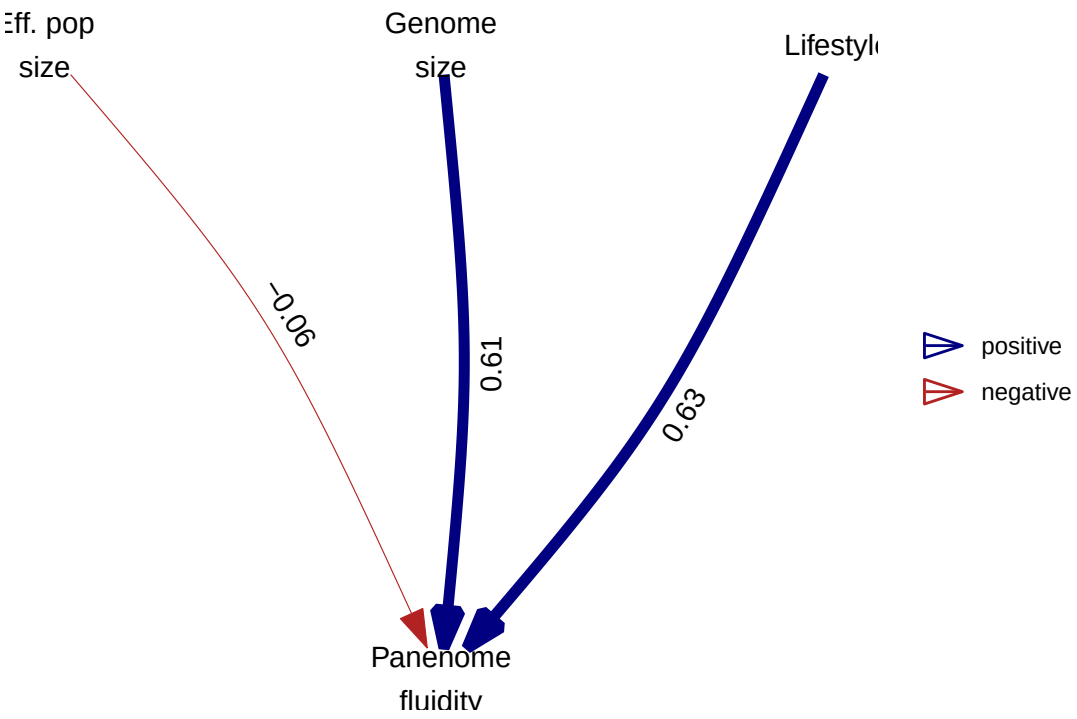
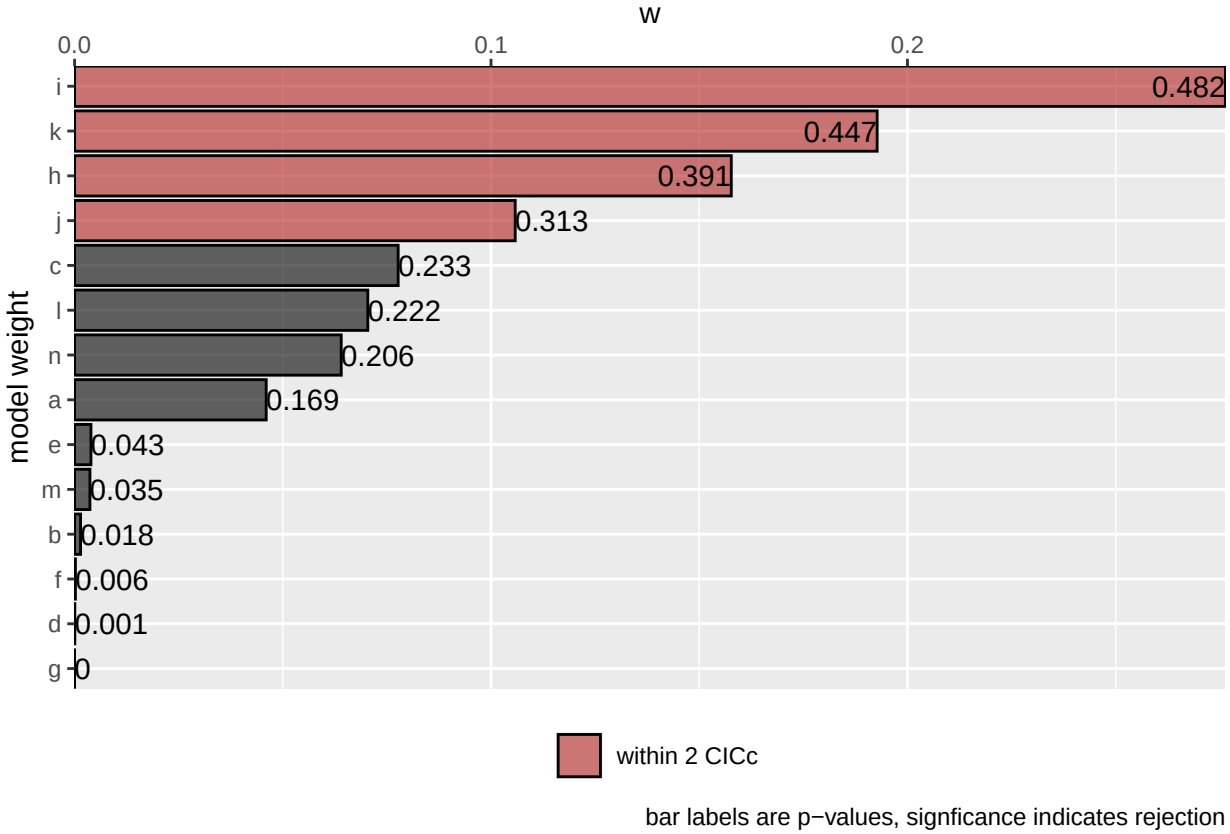


Figure S33: Average of two models with best support, out of a set of simple models.

We then examined a more complex set of models allowing paths between the variables, as in the previous analysis.

We found that four models had similar support.



We then averaged these four models together. This gives the overall best model, shown in the below figure. As with the previous result, it suggests that lifestyle and genome size have a strong direct influence on pangenome fluidity. Although the average best model did include a direct influence of effective on pangenome fluidity, it was a very small coefficient value and also a negative value, in the opposite direction than predicted, suggesting limited support for this path. This is likely because coding the lifestyle traits as binary removes some of the variance, simplifying the variance available to the model. This impacts effective population size because this simplified lifestyle trait appears to have a relatively strong influence on effective population size itself.

Despite these differences, we can make the same qualitative conclusions from this analysis as we did from the previous analysis: lifestyle and genome size directly influence pangenome fluidity, and effective population size is correlated with fluidity potentially because it is itself influenced by lifestyle.

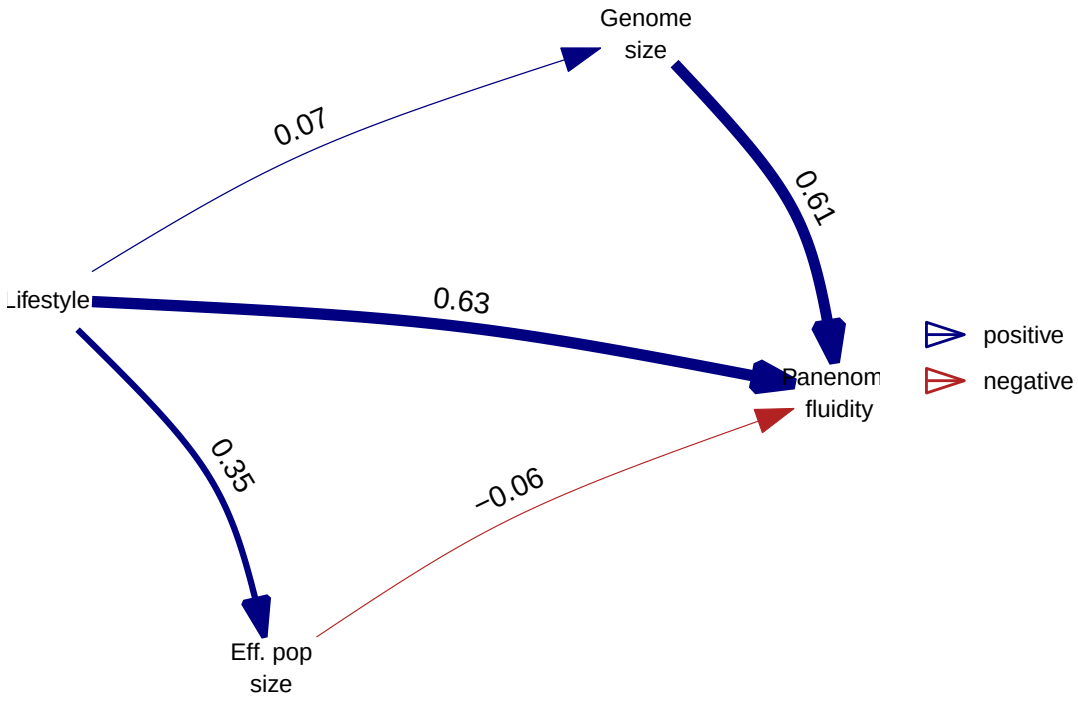


Figure S34: Best model of causal relationships between lifestyle, genome size, effective population size and pangenome fluidity; an average of four models with similar structure.

3.3.2 Multiple component analysis

Next, we used Multiple Component Analysis (MCA) to compute a single variable to capture variation across species' lifestyles.

MCA is a more general form of a Principle Component Analysis (PCA). A PCA plots all the variables in multidimensional space (n dimensions = n variables). It then finds 'components', which are coordinate values that maximise the variation between variables in groups. It does this sequentially, each dimension capturing variation the other(s) hadn't fully captured.

Each variable contributes to these components, but often to different extents – some may contribute virtually nothing to some components.

For our purposes, we ran a MCA to generate a component that explains a high proportion of the variation across species for their combined lifestyle values.

We used MCA rather than PCA because our lifestyle data is best explained as categorical data, and PCA doesn't work so well with this kind of data.

As discussed in the previous section, there are different ways of coding our lifestyle traits depending on whether the intermediate 'both' categories are included for several traits.

We used a similar process to above, where we both: 1. Kept the 'both' in as a category. 2. Combined the 'both' with one of the other two categories.

We then ran an MCA on the data for these two methods of coding the variables to generate a component explanining variation across lifestyle traits. We did this for the 115 species which we had data for all four lifesyle traits. To run the MCA, we used the `MCA()` function from the 'FactoMineR' package. We also used the 'factoextra' package for visualization of the MCA results.

We then conducted analogous phylogenetic path analyses to those described above using this component, across the 75 species which we had both lifestyle and effective population size data for.

The 'scree plot' below shows the initial results from the MCA. Each bar is a dimension which explains different percentages of variance across the species. We can then explore which aspect of variation each dimension explains.

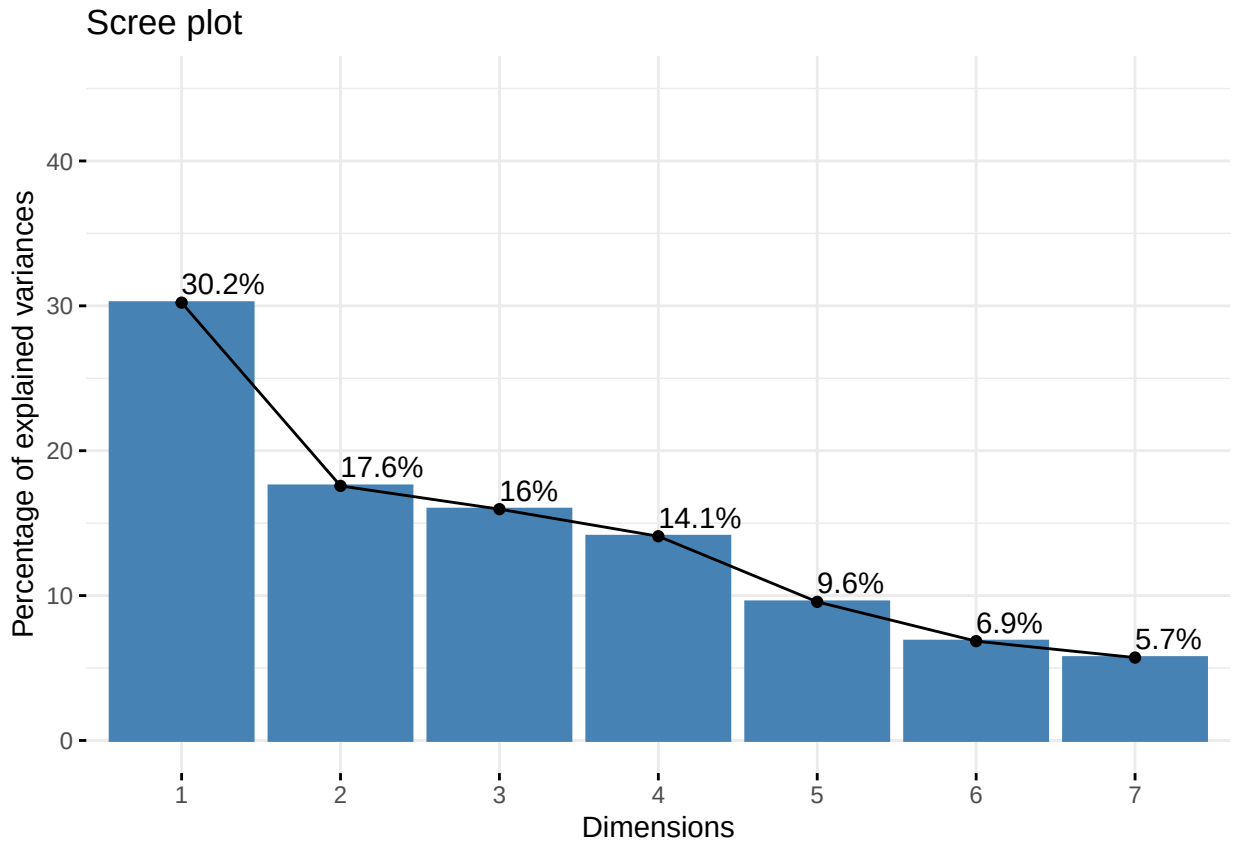


Figure S35: Dimensions produced from MCA analysis on four lifestyle traits across 115 species. Dimensions are ordered from left to right by the percentage of total variation that they explain, which is printed above each bar.

We examined the contribution of the four variables and their categories to the first two dimensions.The first dimension has a more uniform distribution of contributions than the second dimension. All categories contribute, although with considerable variation in quantity, to the first dimension, while a few contribute virtually nothing to the second dimension. A more uniform distribution likely means the dimension captures variation across all the lifestyle traits, which is what we would like to produce from the MCA.

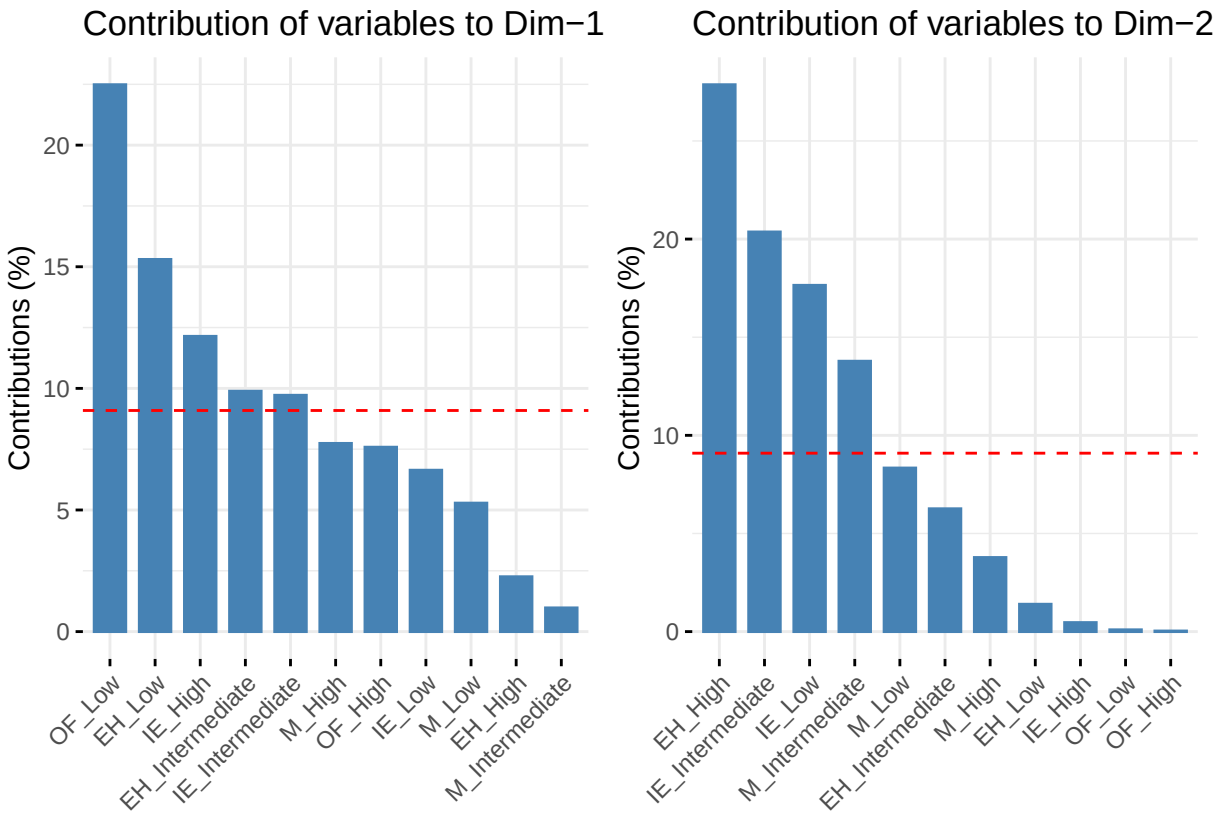


Figure S36: Contribution of variables and their categories to the two dimensions explaining the most variation in lifestyle across the species. OF = host association (obligate/facultative); IE = host location (intra/extracellular); EH = effect on host (pathogen/mutualist); M = motility (non-motile/motile).

Next we examined how the two dimensions correlate with each other, and how this differs between variables.

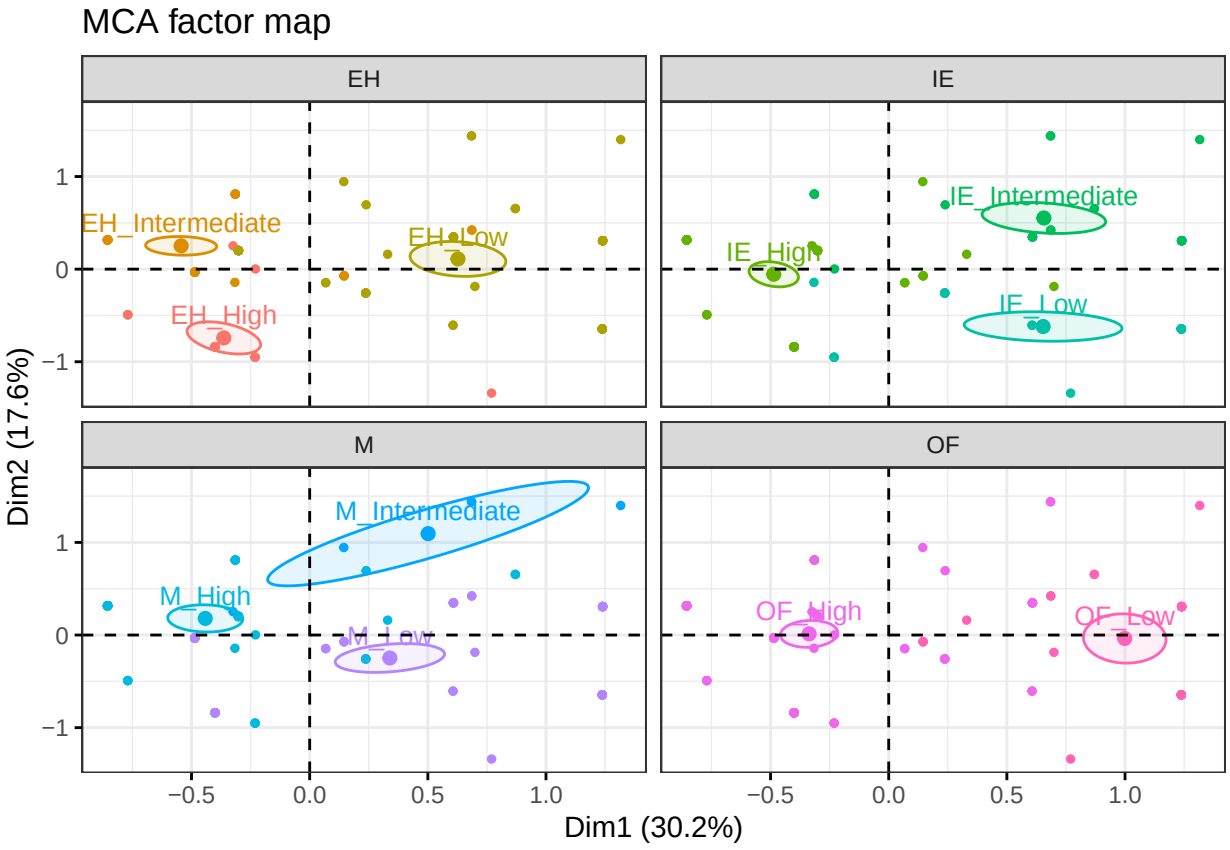


Figure S37: Correlation between dimensions 1 and 2, with a panel for each lifestyle variable highlighting the location of the categories within that lifestyle. Dots represent possible values for a species for each dimension, and colours are to distinguish the categories. Dimension 1 groups species in a way that better reflects what we would like from a single lifestyle variable.

Dimension 1 groups the categories along the axes in the order we would expect. The 'Low' categories (i.e. the ones that correspond to lower lifestyle/environmental variability) are to the right hand side, while the 'High' are to the left for all four categories.

For dimension 1, the 'Intermediate' categories, where present, appear to be similar to one of either the 'High' or 'Low' categories, depending on the lifestyle trait. However, for dimension 2, the 'Intermediate' categories appear more distinct, and for host location and motility in particular, this means the categories do not reflect the gradient of environmental variabilty we would like to capture in a single variable. Furthermore, dimension 2 does not distinguish between obligate and facultative species.

In general, dimension 1 seems to be grouping the categories pretty similarly to the original lifestyle measure. Additionally, it captures almost double the quantity of variation as dimension 2. Consequently, we decided to use dimension 1 going forward for our single measure of bacterial lifestyle.

Dimension 1 is highly correlated with the original lifestyle variable we used in our main analysis.

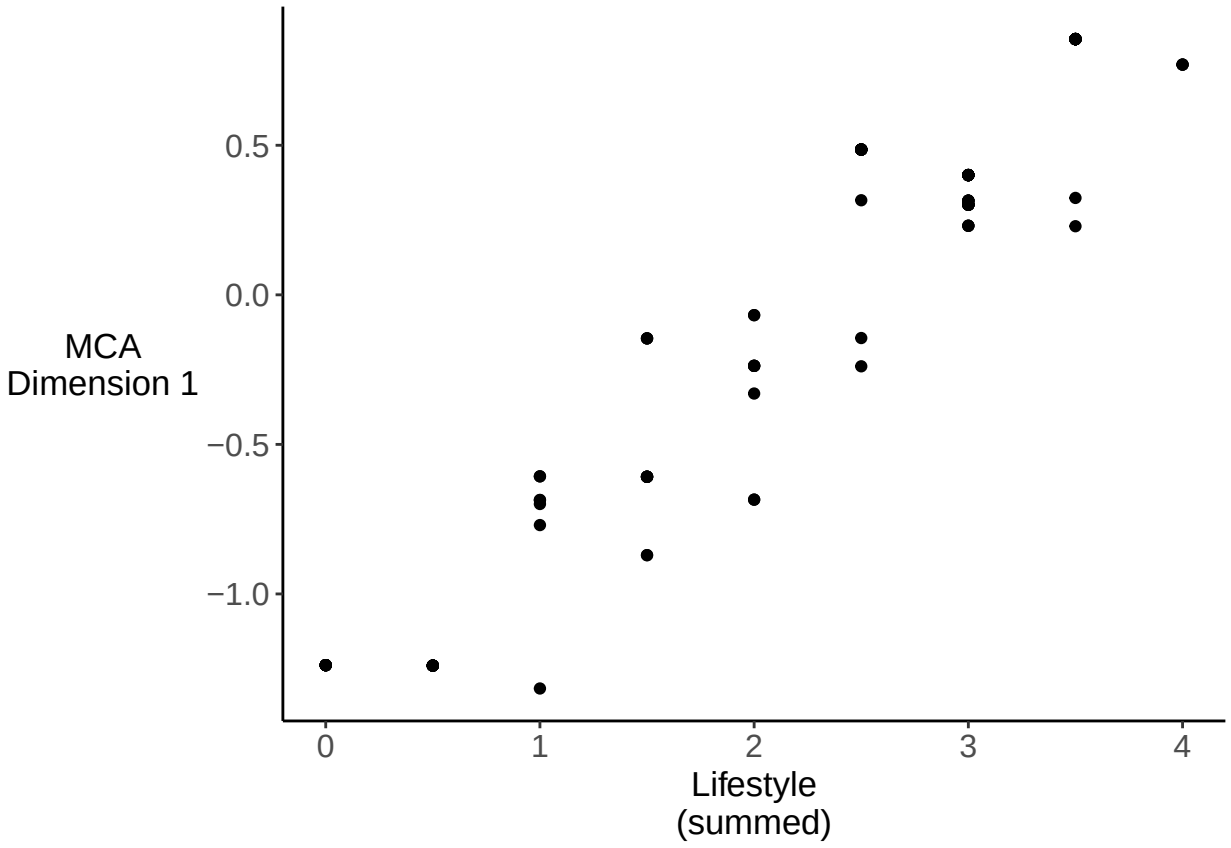


Figure S38: Correlation between our original lifestyle variable used in our main analysis and dimension 1 from the MCA.

We then repeated the path analysis from our main analysis but using dimension 1 instead of our original lifestyle variable.

We first compared a set of simple models, varying by how each of lifestyle, effective population size and genome size influence pangenome fluidity. We found that model c had the highest support; this is the same model that had the highest support in our main analysis.

However, for this analysis, all models were rejected as possible models of causation across the factors (p-value was less than 0.05).

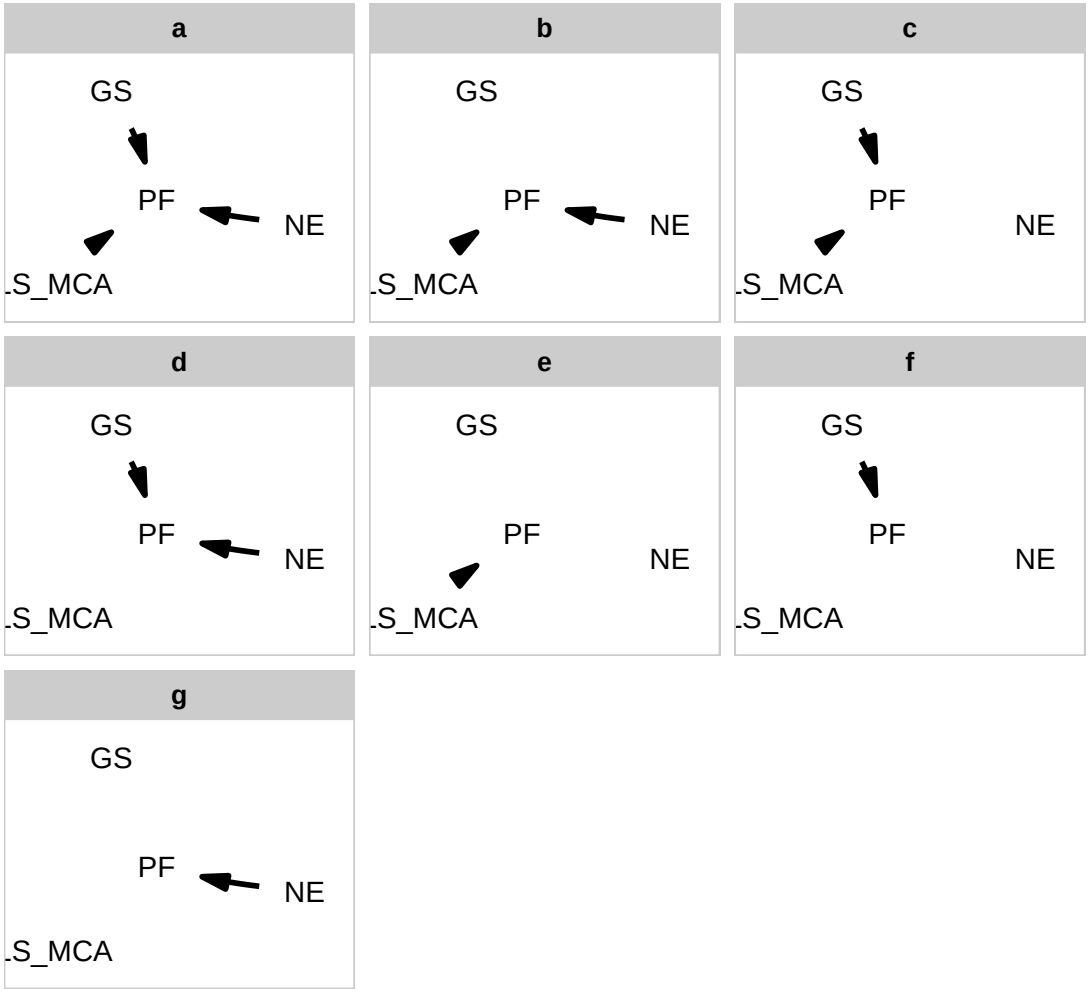


Figure S39: Set of simple models varying by influence of lifestyle (MCA dimension 1), effective population size and genome size, on pangenome fluidity.

Table S52: Comparison of support for a simple set of models varying by which of lifestyle (MCA dimension 1), effective population size and genome size causes pangenome fluidity.

	w	p	CICc
c	0.4614	0.0467	28.9482
a	0.3861	0.0340	29.3044
b	0.1061	0.0168	31.8870
e	0.0458	0.0119	33.5673
f	0.0005	0.0004	42.6197
d	0.0001	0.0001	45.6708
g	0.0000	0.0000	57.2711

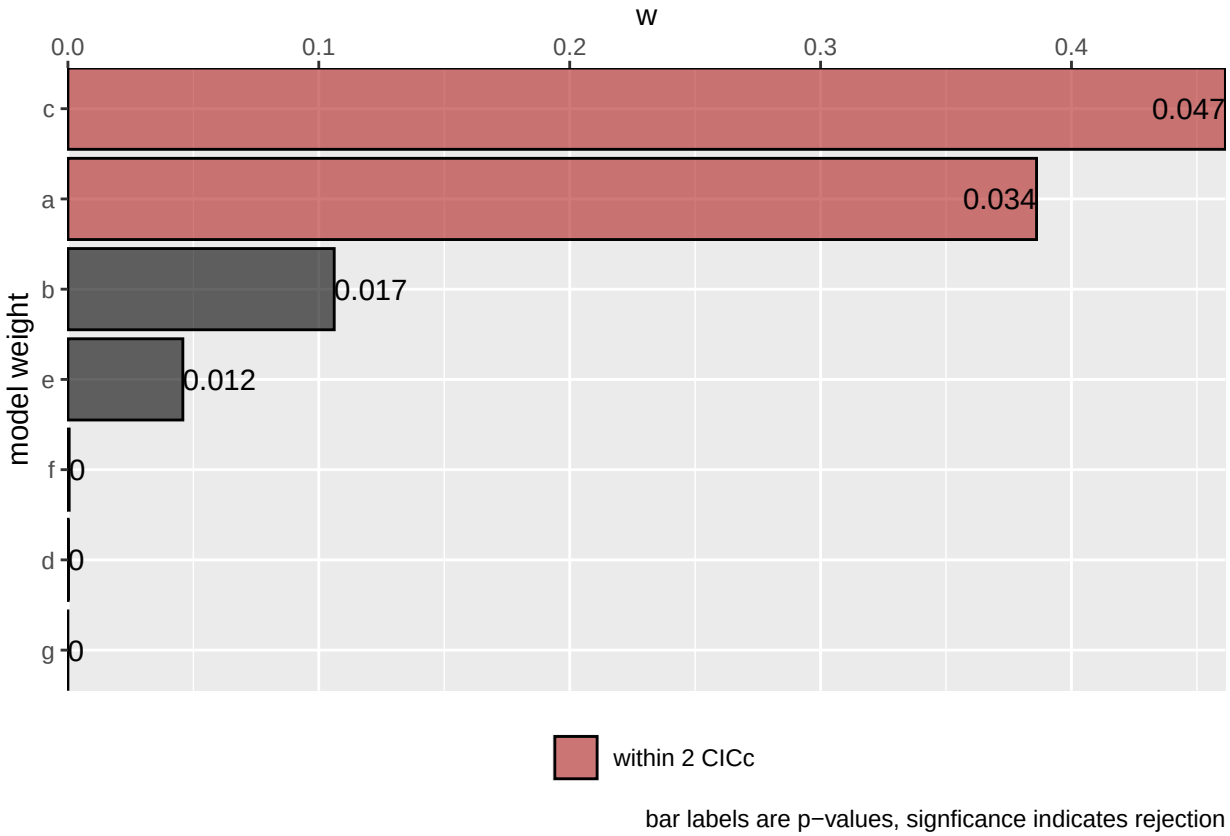


Figure S40: Comaparison of support for a simple set of models varying by which of lifestyle (MCA dimension 1), effective population size and genome size causes pangenome fluidity.

We examined the specific results of the model with the highest support, model c. The model was rejected because the lifestyle MCA variable was not conditionally independent from genome size. Thus, a more complex model is required to fully explain any causal relationships between these variables.

Table S53: Specific resuls for model c, showing support for each conditional independency specified by the model. Lifestyle and genome size are not conditionally independent, indicated by the p-value of less than 0.05, meaning they are significantly correlated.

d_sep	p
GS ~ NE	0.2503
LS_MCA ~ NE	0.1140
PF ~ GS + LS_MCA + NE	0.3534
LS_MCA ~ GS	0.0384

We next compared support for a set of more complex models, allowing paths between lifestyle (MCA dimension 1), genome size and effective population size.

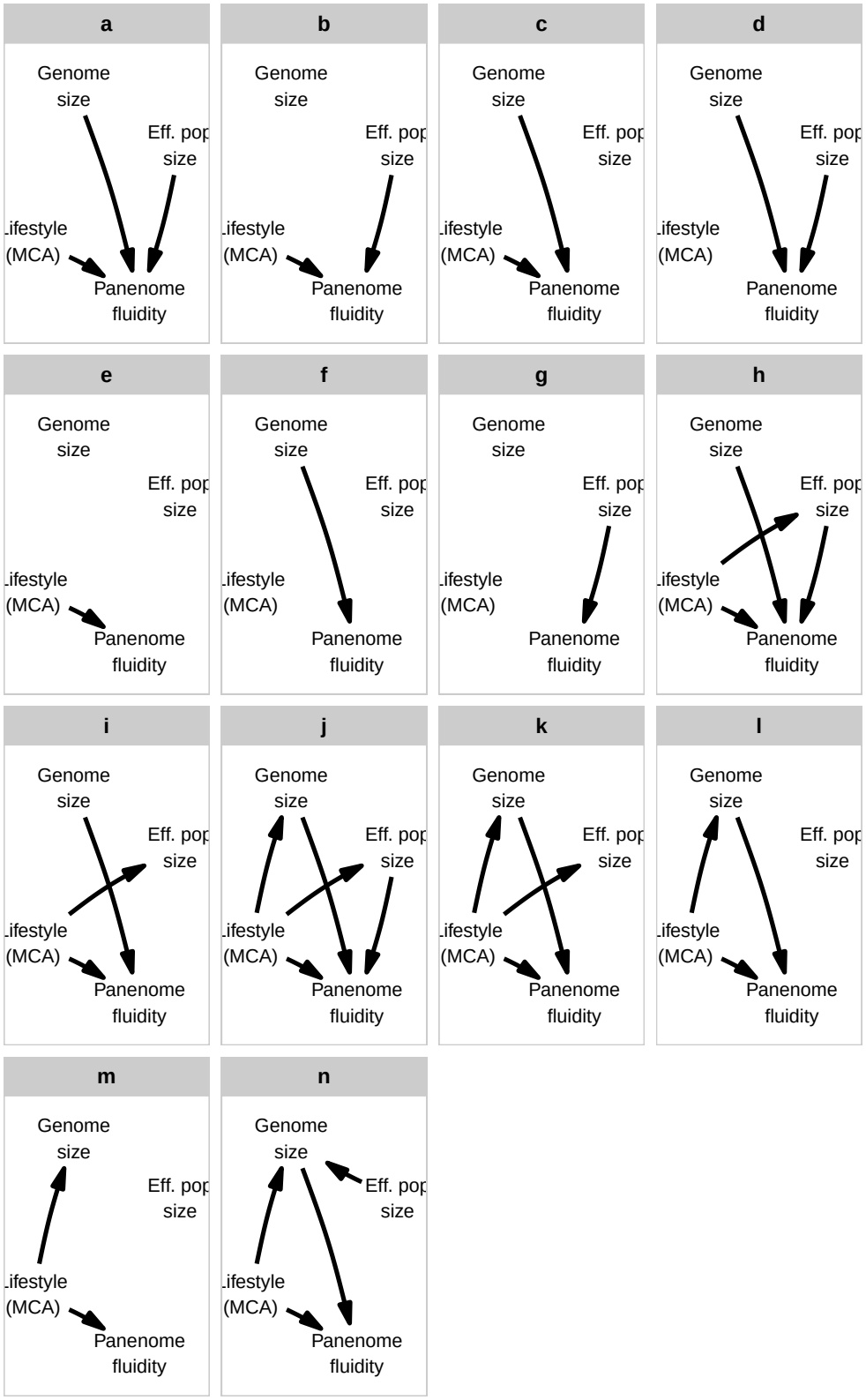


Figure S41: A set of more complex causal models, varying by which of lifestyle (MCA dimension 1), genome size and effective population size directly influence pangenome fluidity, and how they might influence each other.

We found that the model with the highest support was model n, which was also the case for our main analysis.

Table S54: Comparison of set of more complex causal models, varying by which of lifestyle (MCA dimension 1), genome size and effective population size directly influence pangenome fluidity, and how they might cause each other.

	w	p	CICc
n	0.2753	0.1697	24.6047
k	0.2669	0.1658	24.6666
j	0.2071	0.1105	25.1738
l	0.1068	0.0939	26.4991
i	0.0359	0.0430	28.6764
c	0.0314	0.0467	28.9482
h	0.0290	0.0274	29.1064
a	0.0263	0.0340	29.3044
m	0.0110	0.0227	31.0476
b	0.0072	0.0168	31.8870
e	0.0031	0.0119	33.5673
f	0.0000	0.0004	42.6197
d	0.0000	0.0001	45.6708
g	0.0000	0.0000	57.2711

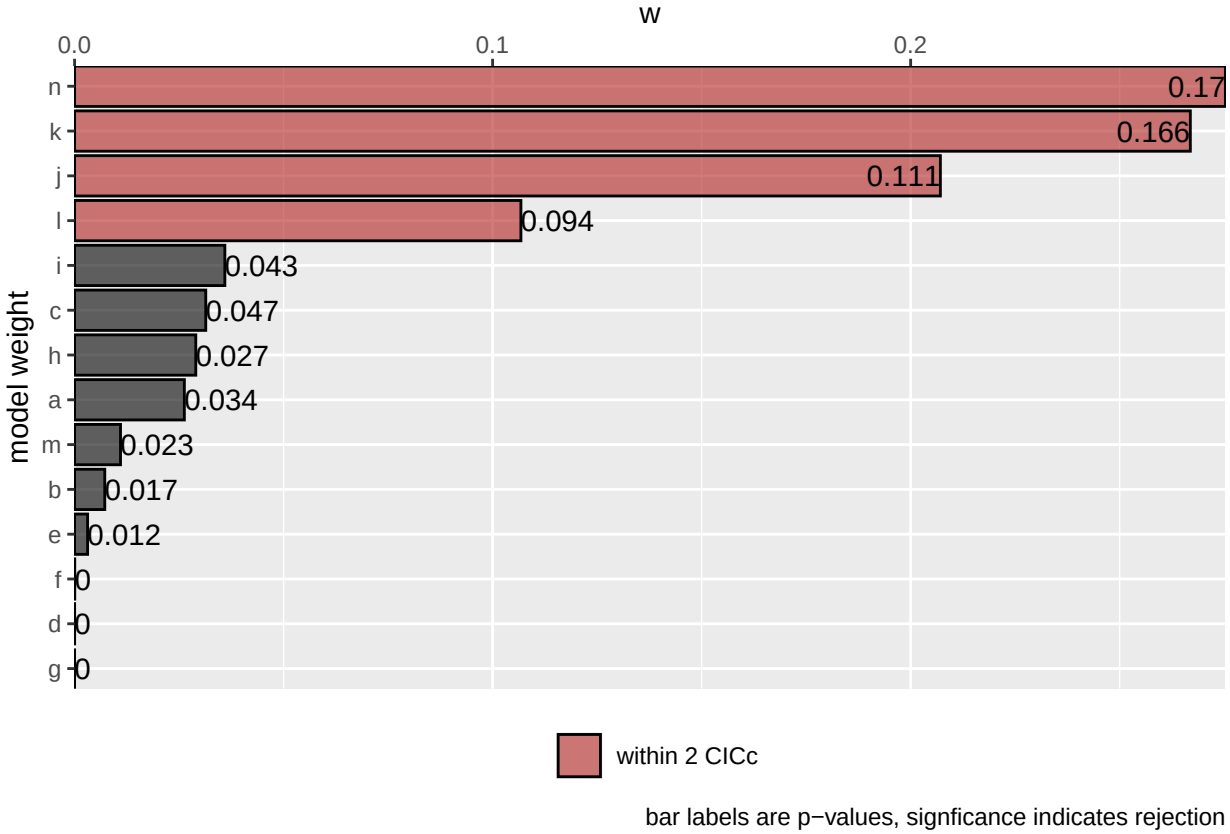


Figure S42: Comparison of support for a more complex set of models varying by which of lifestyle (MCA dimension 1), genome size and effective population size directly influence pangenome fluidity, and how they might cause each other.

Additionally, we found that three other models had support within 2 CICs of model n. We can average these four models together, either by averaging the values of each path only when they are present in a model (first figure below), or by setting that path to 0 when it is not present (second figure below).

Compared to our main analysis, the causal structure between the variables appears less well resolved. This is likely because dimension 1 of the MCA explained only 30% of the variance across species, and did not neccesarily order the categories of each lifestyle trait in a manner which best reflects the biology underlying that trait, particularly with relation to how it might influence environmental variability.

However, the result is generally the same as our main analysis: both genome size and lifestyle have a direct influence on pangenome fluidity, with little evidence of a direct influence of effective population size on pangenome fluidity (this path was present in only one of the four models, and when it was present the correlation coefficient was low).

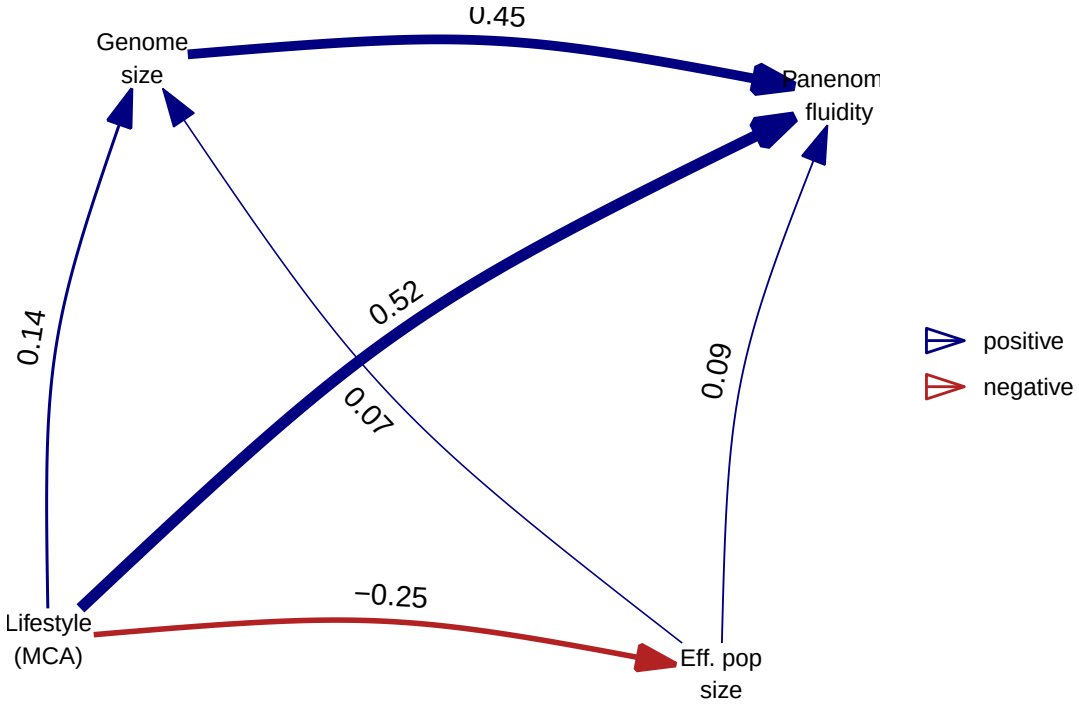


Figure S43: Best model of causal relationships between lifestyle, genome size, effective population size and pangenome fluidity; an average of four models.

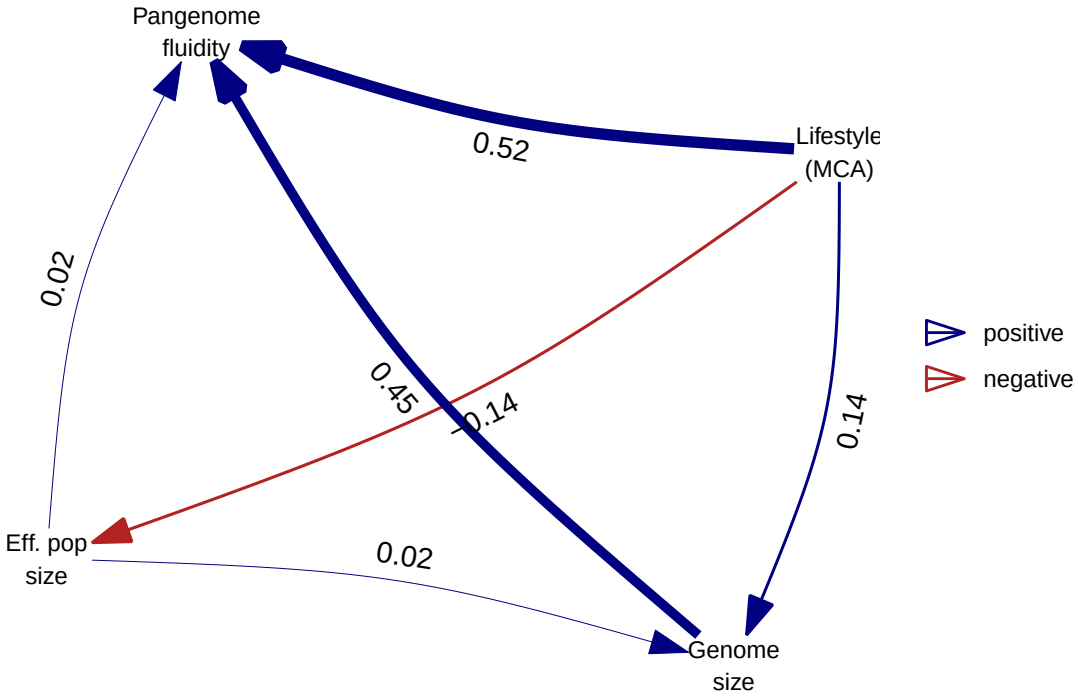


Figure S44: Best model of causal relationships between lifestyle, genome size, effective population size and pangenome fluidity; an average of four models with similar structure, with correlation coefficients weighted by setting absent paths to zero when averaging models.

3.4 Host-association and effective population size

We also ran very simple phylogenetic path analsyis, considering only two factors' potential influence on species' pangenome fluidity: whether a species was host-associated or free-living, and species' effective population size.

We used our categorization of species into four groups based on the frequency of their host association, as in Figure 2a in the main text and Figure 3 of this document. These categories were: Host-associated, mostly host, mostly free, and free-living. Because we only considered species which were at least sometimes host-associated in previous analyses, we did not include the four 'free-living' species. Of those remaining species, we coded host-associated as 0, mostly host as 0.5, and mostly free-living as 1, to convert this into a numeric variable.

Using this variable, we used phylogenetic path analysis to compare support for a very simple set of models.

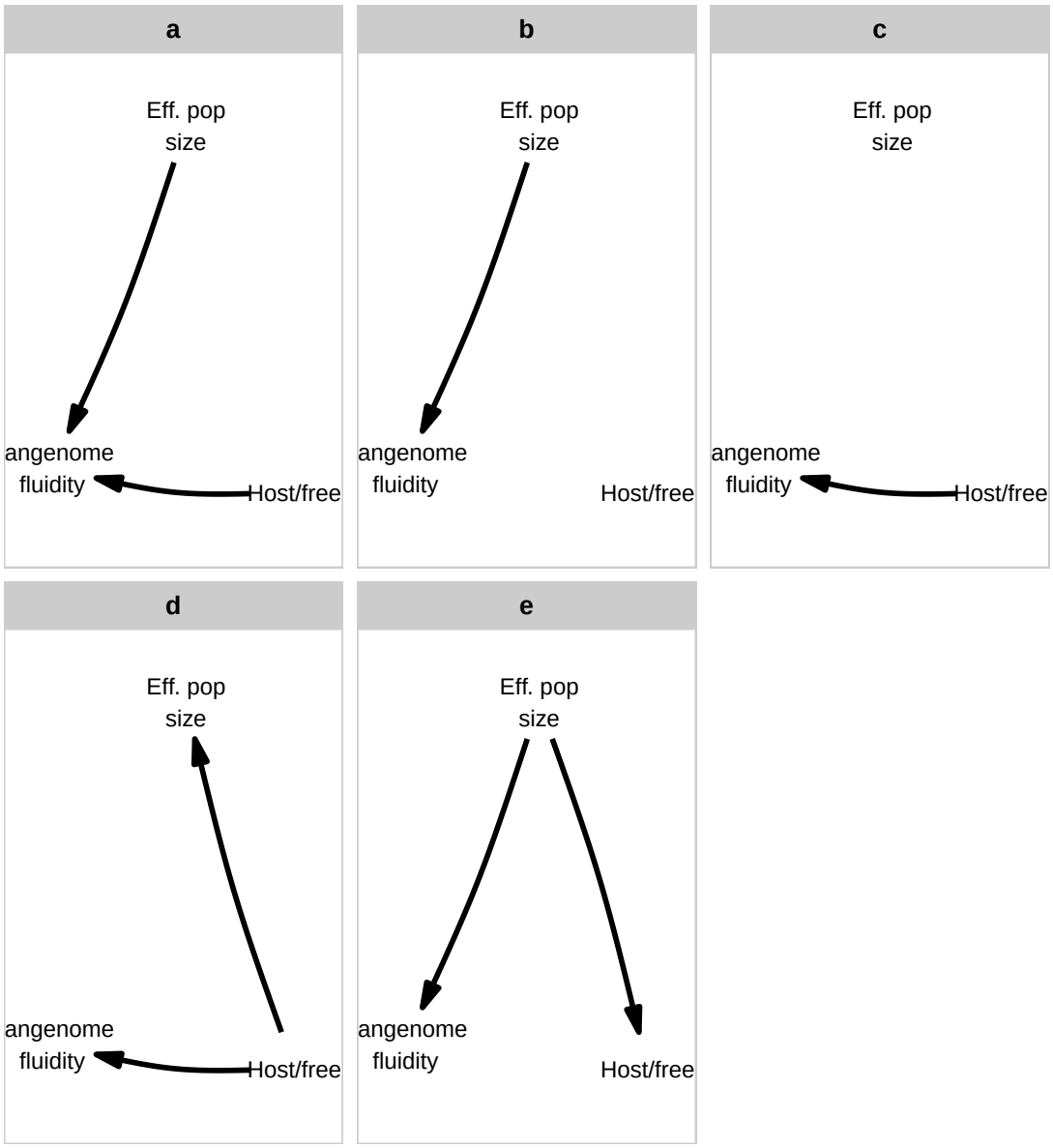


Figure S45: Set of simple models varying by which of effective population size and whether a species is mostly host-associated or free-living influences pangenome fluidity.

Three of the five models were not rejected and were within 2 CICs of each other. The remaining two models (model b and model e) were rejected (their p-values were less than 0.01). Model b had only an influence of effective population size on pangenome fluidity, but no influence of lifestyle. Model e was the same as model b, with only a direct influence of effective population size on pangenome fluidity, but with an additional influence of effective population size on lifestyle. These were both rejected, suggesting no support for effective population size being the key factor driving pangenome fluidity compared to lifesyle (in this

case, how frequently a species' is host-associated).

Table S55: Comparison of set of more complex causal models, varying by which of host-association (host, mostly host, mostly free) and effective population size influence pangenome fluidity.

	w	p	CICc
c	0.3797	0.4477	12.2746
d	0.3611	0.4710	12.3754
a	0.2555	0.3333	13.0669
b	0.0019	0.0064	22.8560
e	0.0018	0.0024	22.9568

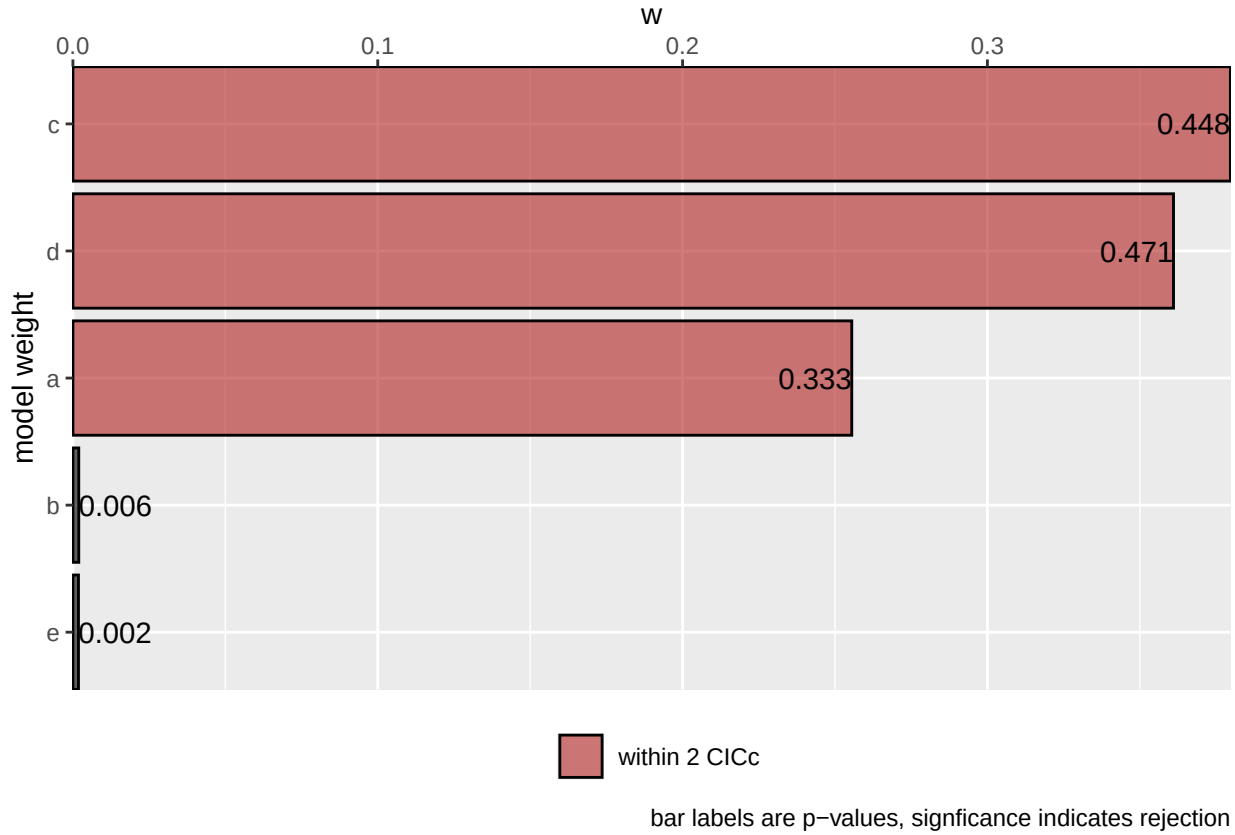


Figure S46: Comparison of set of more complex causal models, varying by which of host-association (host, mostly host, mostly free) and effective population size influence pangenome fluidity.

We averaged the three models with similar support together, both averaging only when that path is present (first plot below) and setting the path to zero when it is not present (second plot below). In summary, these models suggest strong support for a causal influence of lifestyle, here simply the extent to which a species is host-associated, on pangenome fluidity. We find some evidence that lifestyle also influences effective population size. Finally, we find no evidence that effective population size has a key direct influence on pangenome fluidity.

These results agree with the best model in our main analysis, when we considered more lifestyle traits, and also genome size.

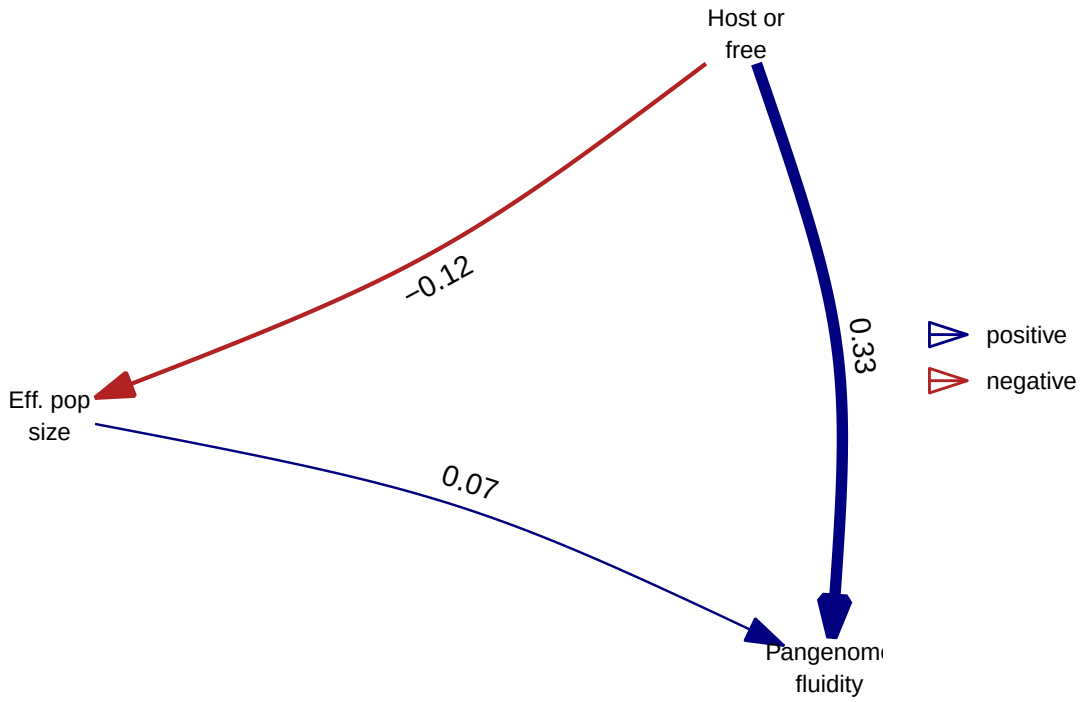


Figure S47: Average best model of causal relationships between host-association (host-associated, mostly host and mostly free-living), effective population size and pangenome fluidity.

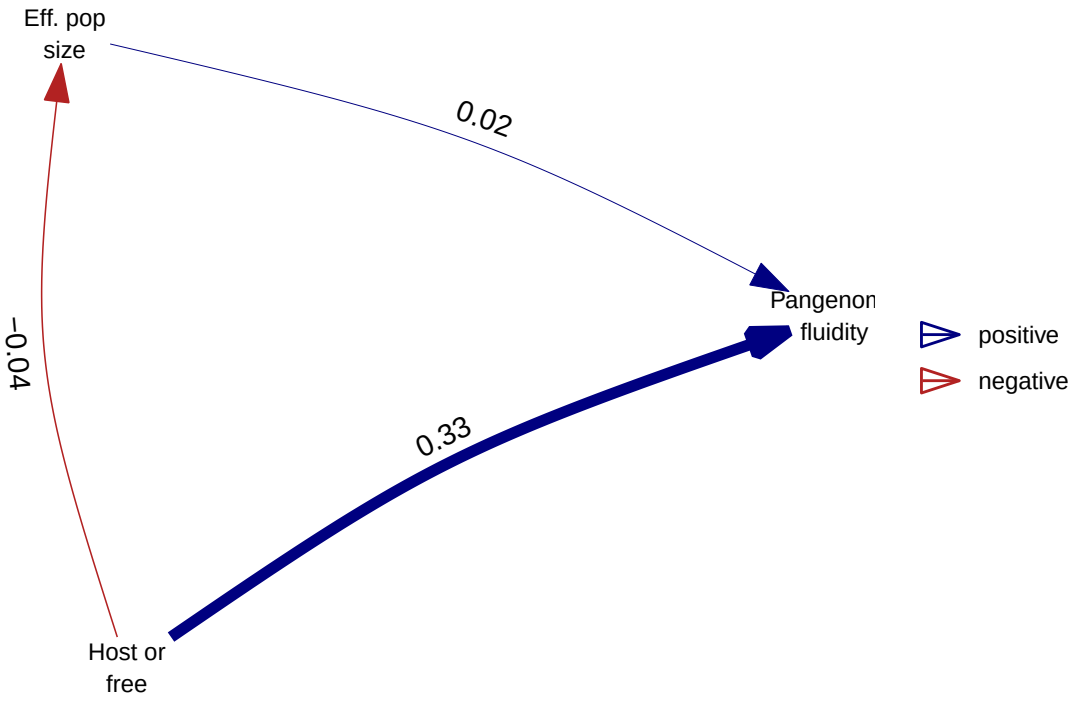


Figure S48: Average best model of causal relationships between host-association (host-associated, mostly host and mostly free-living), effective population size and pangenome fluidity; an average of four models with similar structure, with correlation coefficients weighted by setting absent paths to zero when averaging models.

4 Other measures of pangenome variation across species

We also identified 'core' and 'accessory' genes, to examine the extent to which the relative proportion of these correlated with our main measure of pangenome structure, pangenome fluidity. There is some inconsistency in how to define core genes in the literature. Some authors define core genes as those present in all genomes, while others choose a high threshold that allows for an occasional genome not to carry that gene.

Accordingly, we defined core genes using three different thresholds: (1) genes present in 100% of genomes; (2) genes present in $\geq$ 90% of genomes; (3) genes that present in $\geq$ 80% of genomes. We also identified genes that only existed in a small subset of genomes, which we refer to here as 'accessory' genes for simplicity. We defined these accessory genes with two thresholds: (1) genes present in $\leq$ 10% of genomes; (2) genes present in $\leq$ 20% of genomes.

We then calculated the percentage of core and accessory genes for each species' pangenome. We did this for all thresholds of core and accessory genes stated above.

In addition, we were also interested in understanding the average percentage of core genes at the individual genome level, rather than the entire pangenome level. Consequently, we calculated the proportion of core genes at the genome level for each species as followed:

$$Proportion\ of\ core\ genes\ in\ genome = \sum_i^n \frac{\frac{number\ of\ core\ genes\ in\ genome\ i}{number\ of\ genes\ in\ genome\ i}}{number\ of\ genomes(n)}$$

We did this for the three thresholds of core genes, as stated above.

Some genomes did not contain any genes found in ≤ 10 or $\leq 20\%$ and so this would bias our calculations of the average percentage of accessory genes at the genome level. Therefore, we only looked at the percentage of accessory genes at the pangenome level.

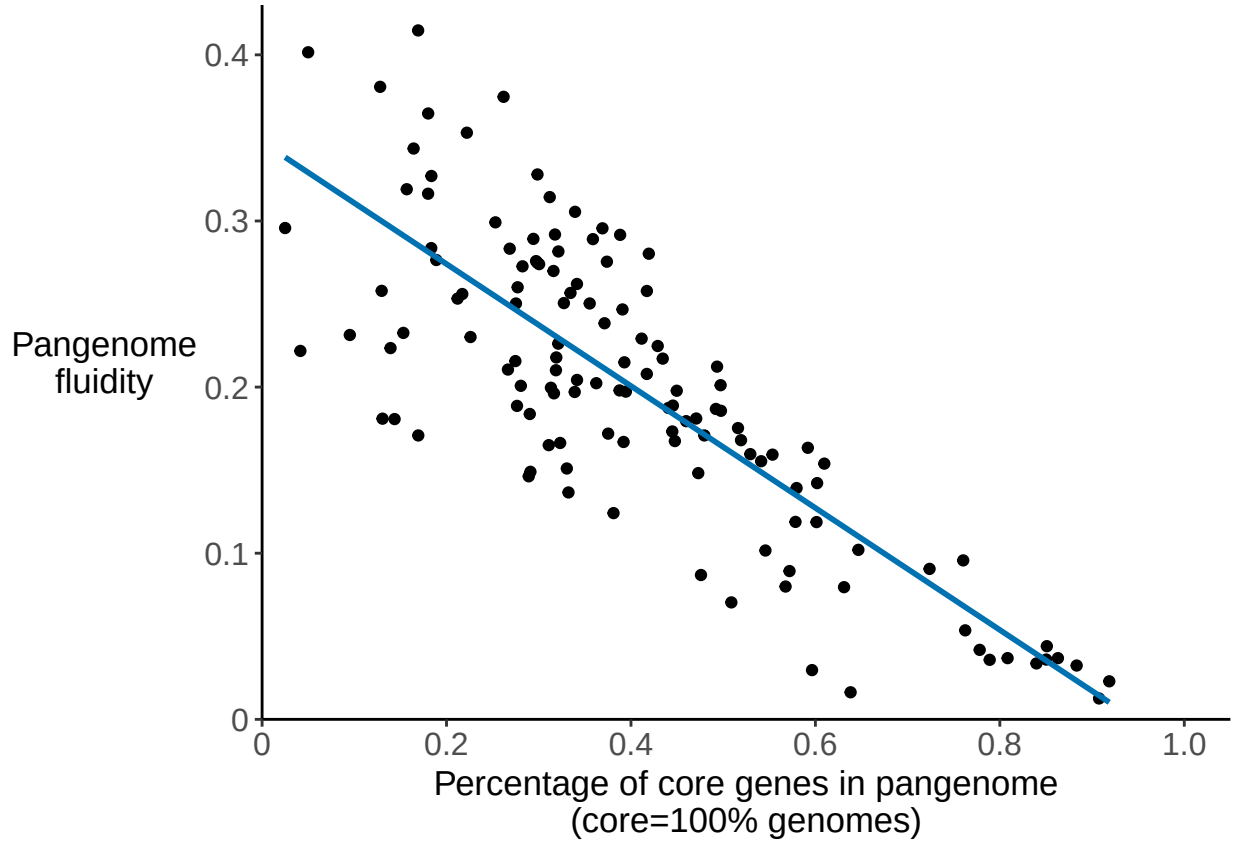


Figure S49: Pangenome fluidity is highly correlated with the proportion of core genes in the pangenome. Scatterplot showing how pangenome fluidity varies with the percentage of genes in a species' pangenome which are 'core', defined in this plot as present in 100% of genomes. Data includes all 126 species.

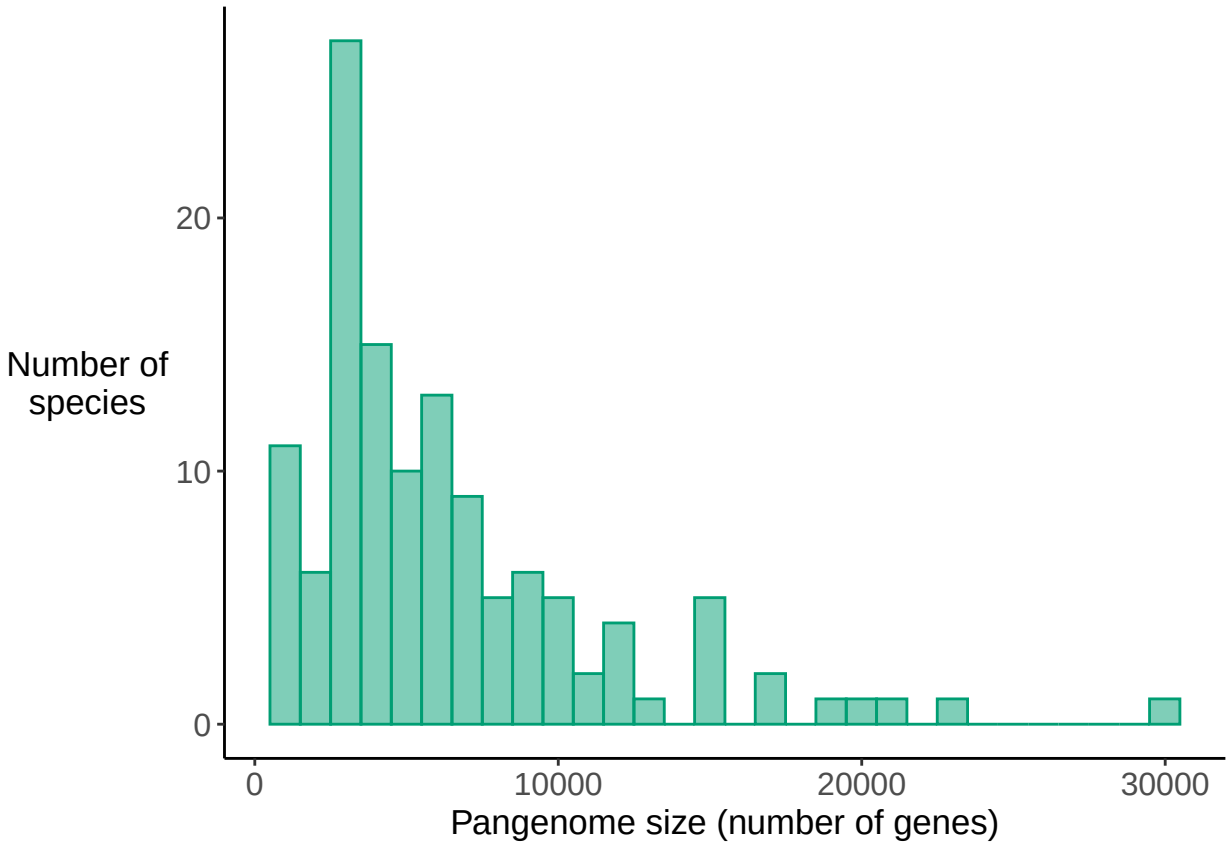


Figure S50: Pangenome size varies considerably across species. Histogram showing the variation in pangenome size, meaning the number of unique genes sequenced across all genomes of a species. Data is for all 126 species.

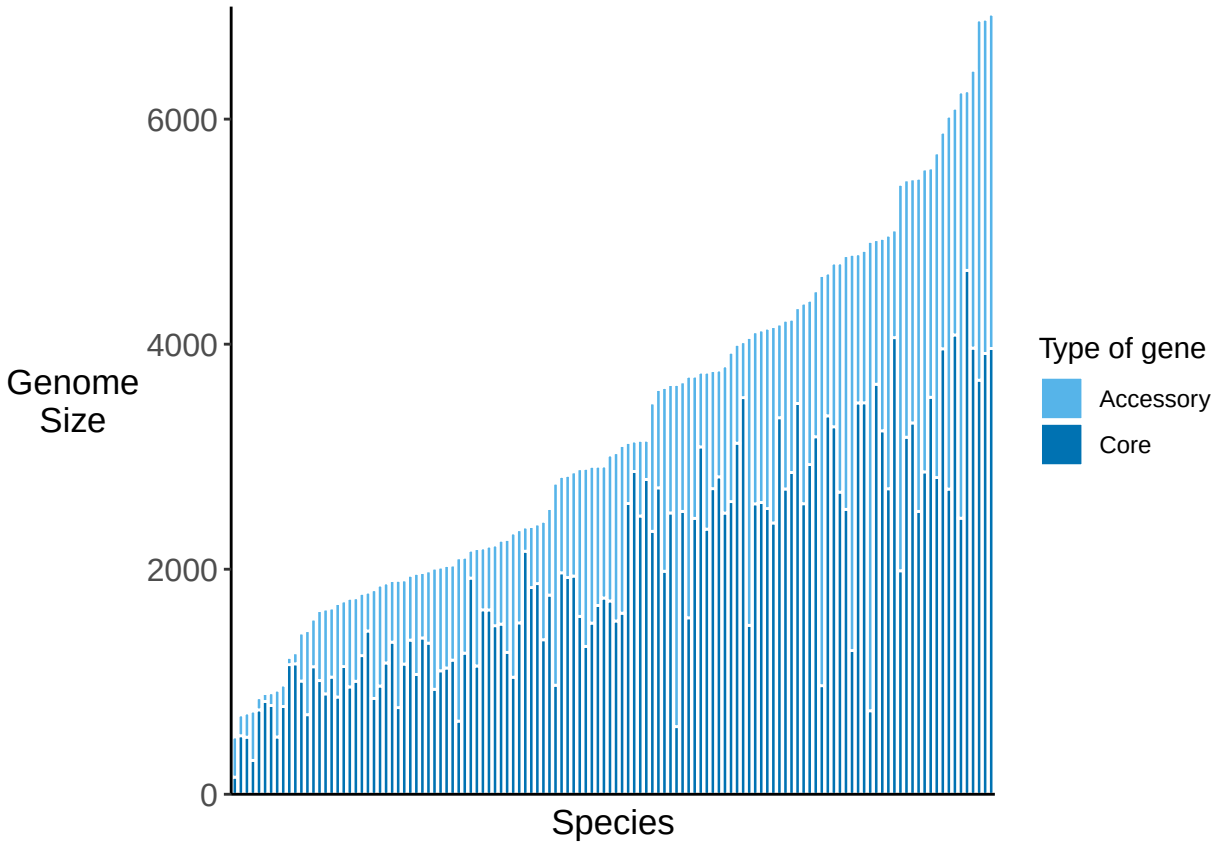


Figure S51: A stacked bar plot, with a bar for each species; the total height is the average genome size for that species, the 'core' are the genes present in 100% of genomes, and the 'accessory' genes are any remaining genes not present in 100% of genomes.

Pangenome fluidity was not influenced by the number of genomes we had for each species.

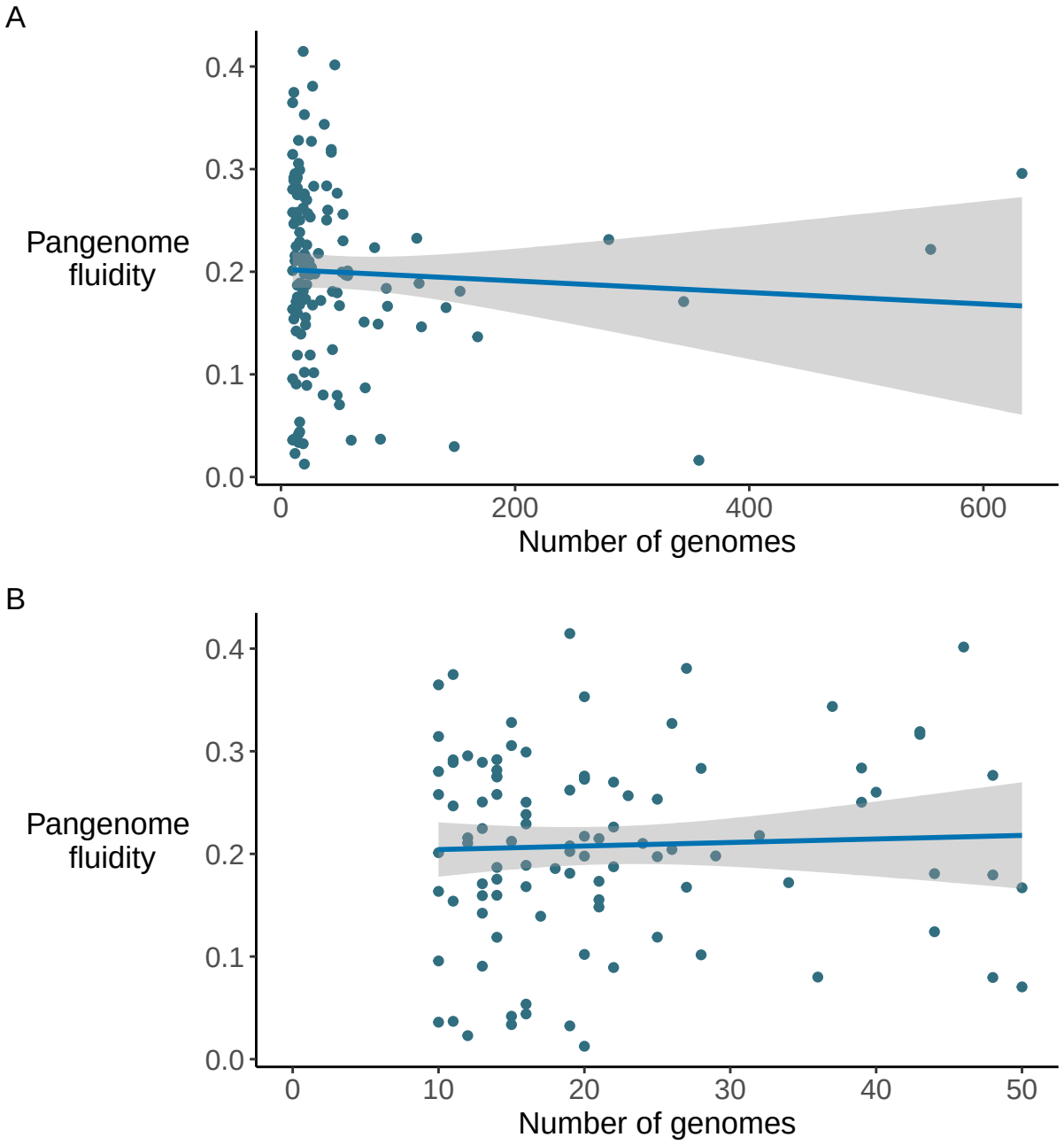


Figure S52: No correlation between pangenome fluidity and the number of genomes per species. A. All species; B. Subset with species between 10-50 genomes.

In contrast, the percentage of core genes in the pangenome and the size of the pangenome were both correlated with the number of genomes we had for each species.

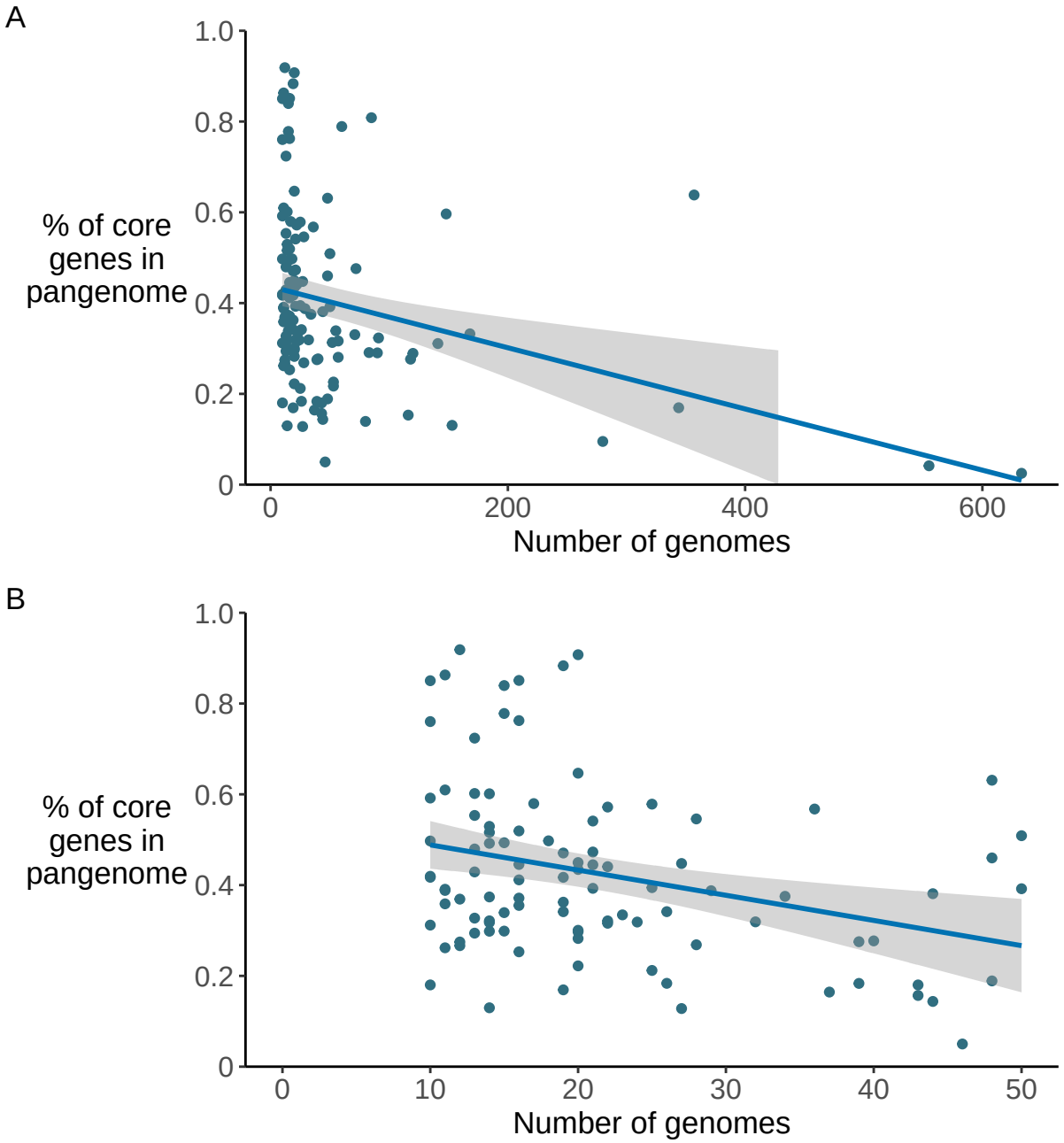


Figure S53: The percentage of the pangenome which is core genes is negatively correlated with the number of genomes per species. A. All species; B. Subset with species between 10-50 genomes.

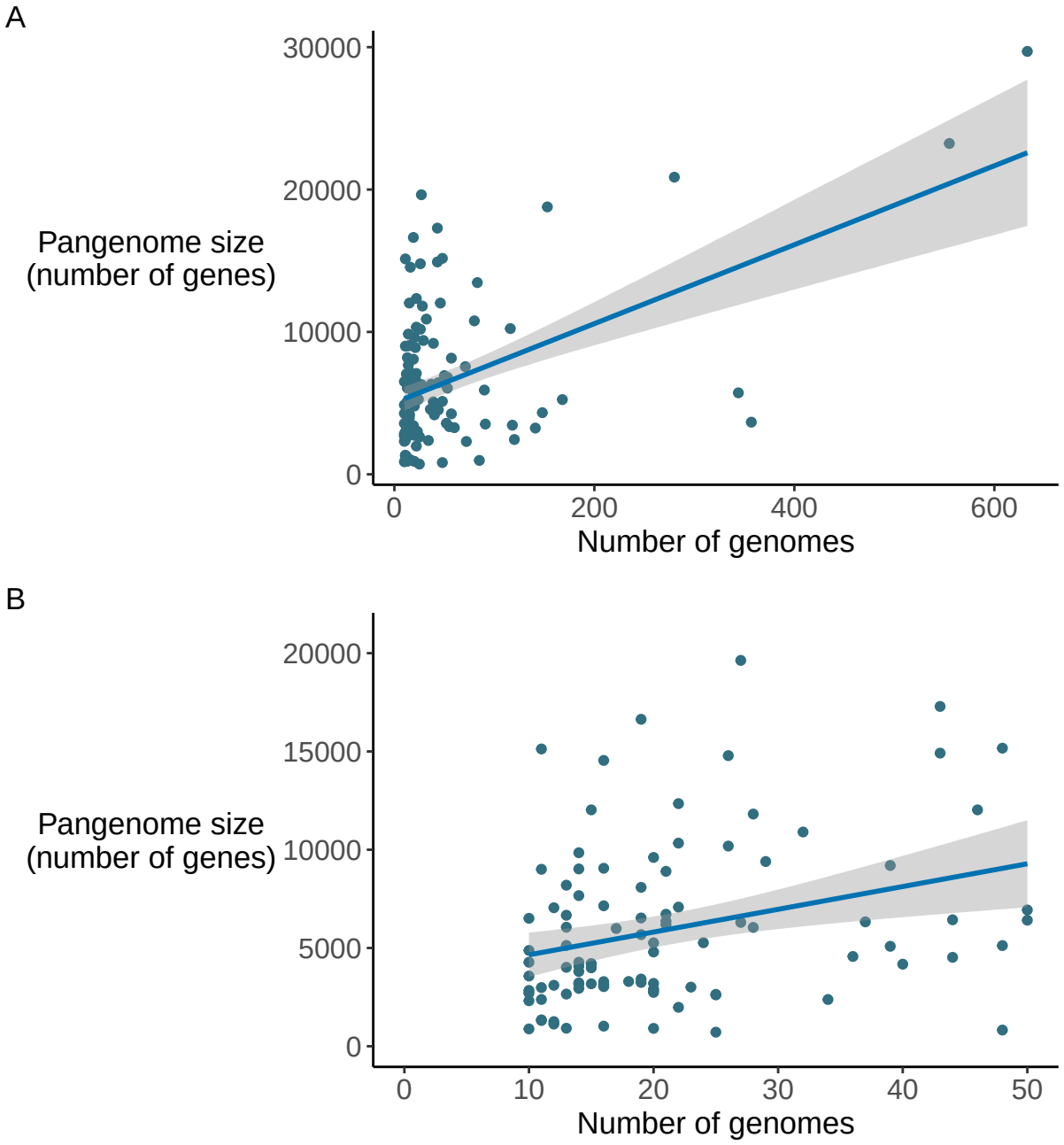


Figure S54: The size of the pangenome is positively correlated with the number of genomes per species. A. All species; B. Subset with species between 10-50 genomes.

5 Phylogeny

Below is the phylogeny we used for all analyses in our paper.

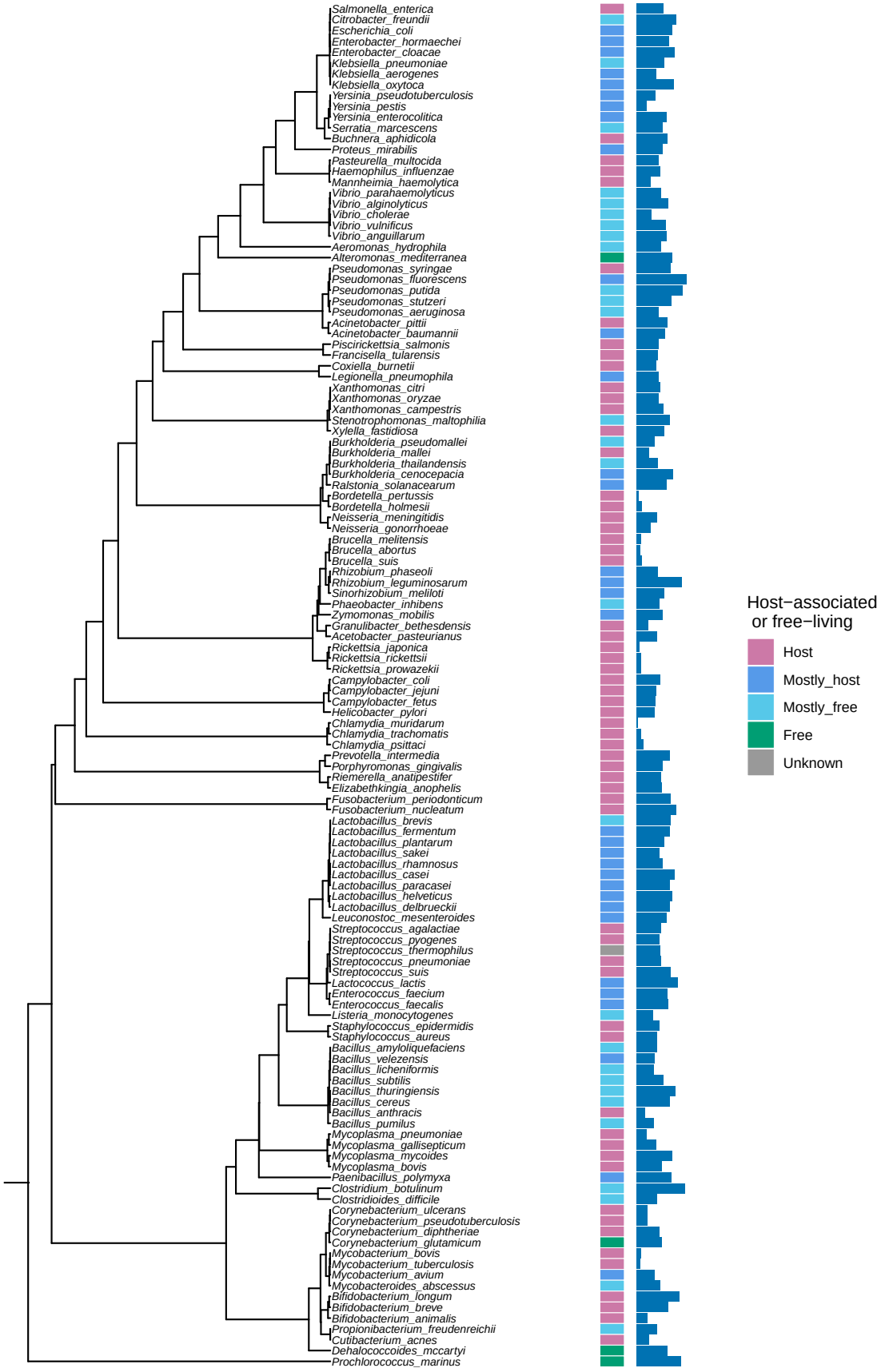


Figure S55: Phylogeny of all 126 species in our dataset, which is available to download in ultrametric nexus format from github (*link*). Coloured squares indicate whether a species is host-associated, free-living, or a mixture of both, and bars indicate pangenome fluidity.

6 Species lifestyle table

Below is a table with all our species and their categories for the key lifestyle variables we examined.

Table S56: Information for key lifestyle traits for all 126 species in our dataset.

Species	Host or free	Host association	Host location	Effect on host	Motility
Acetobacter_pasteurianus	Host	Facultative	Extracellular	Mutualist	Both
Acinetobacter_baumannii	Mostly_host	Facultative	Extracellular	Pathogen	Both
Acinetobacter_pittii	Host	Facultative	Extracellular	Pathogen	Non-motile
Aeromonas_hydrophila	Mostly_free	Facultative	Extracellular	Pathogen	Motile
Alteromonas_mediterranea	Free				Motile
Bacillus_amyloliquefaciens	Mostly_free	Facultative	Extracellular	Mutualist	Motile
Bacillus_anthracis	Host	Obligate	Both	Pathogen	Non-motile
Bacillus_cereus	Mostly_free	Facultative	Extracellular	Both	Motile
Bacillus_licheniformis	Mostly_free	Facultative	Extracellular	Both	Motile
Bacillus_pumilus	Mostly_free	Facultative	Extracellular	Both	Motile
Bacillus_subtilis	Mostly_free	Facultative	Extracellular	Mutualist	Motile
Bacillus_thuringiensis	Mostly_free	Facultative	Extracellular	Pathogen	Motile
Bacillus_velezensis	Mostly_host	Facultative	Extracellular	Mutualist	Motile
Bifidobacterium_animalis	Host	Facultative	Extracellular	Mutualist	Non-motile
Bifidobacterium_breve	Host	Facultative	Extracellular	Mutualist	Non-motile
Bifidobacterium_longum	Host	Facultative	Extracellular	Mutualist	Non-motile
Bordetella_holmesii	Host	Obligate	Both	Pathogen	Non-motile
Bordetella_pertussis	Host	Obligate	Both	Pathogen	Both
Brucella_abortus	Host	Facultative	Both	Pathogen	Non-motile
Brucella_melitensis	Host	Facultative	Both	Pathogen	Non-motile
Brucella_suis	Host	Facultative	Both	Pathogen	Non-motile
Buchnera_aphidicola	Host	Obligate	Intracellular	Mutualist	Non-motile
Burkholderia_cenocepacia	Mostly_host	Facultative	Intracellular	Both	Motile
Burkholderia_mallei	Host	Obligate	Intracellular	Pathogen	Non-motile
Burkholderia_pseudomallei	Mostly_free	Facultative	Intracellular	Pathogen	Motile
Burkholderia_thailandensis	Mostly_free	Facultative	Intracellular	Pathogen	Motile
Campylobacter_coli	Host	Facultative	Both	Both	Motile
Campylobacter_fetus	Host	Facultative	Both	Both	Motile
Campylobacter_jejuni	Host	Facultative	Both	Both	Motile
Chlamydia_muridarum	Host	Obligate	Intracellular	Pathogen	Non-motile
Chlamydia_psittaci	Host	Obligate	Intracellular	Pathogen	Non-motile
Chlamydia_trachomatis	Host	Obligate	Intracellular	Pathogen	Non-motile
Citrobacter_freundii	Mostly_free	Facultative	Extracellular	Both	Motile
Clostridioides_difficile	Mostly_free	Facultative	Extracellular	Both	Motile
Clostridium_botulinum	Mostly_free	Facultative	Extracellular	Pathogen	Motile
Corynebacterium_diphtheriae	Host	Obligate	Both	Pathogen	Non-motile
Corynebacterium_glutamicum	Free				Non-motile
Corynebacterium_pseudotuberculosis	Host	Obligate	Intracellular	Pathogen	Non-motile
Corynebacterium_ulcerans	Host	Obligate	Both	Pathogen	Non-motile
Coxiella_burnetii	Host	Obligate	Intracellular	Pathogen	Non-motile
Cutibacterium_acnes	Host	Obligate	Extracellular	Both	Non-motile
Dehalococcoides_mccartyi	Free				Non-motile
Elizabethkingia_anophelis	Host	Unknown	Extracellular	Unknown	Non-motile
Enterobacter_cloacae	Mostly_host	Facultative	Extracellular	Both	Motile
Enterobacter_hormaechei	Mostly_host	Facultative	Extracellular	Both	Motile
Enterococcus_faecalis	Mostly_host	Facultative	Extracellular	Both	Motile
Enterococcus_faecium	Mostly_host	Facultative	Extracellular	Both	Non-motile
Escherichia_coli	Mostly_host	Facultative	Extracellular	Both	Motile
Francisella_tularensis	Host	Facultative	Intracellular	Pathogen	Non-motile
Fusobacterium_nucleatum	Host	Obligate	Extracellular	Both	Non-motile
Fusobacterium_periodonticum	Host	Obligate	Extracellular	Both	Non-motile
Granulibacter_bethesdensis	Host	Unknown	Intracellular	Pathogen	Non-motile
Haemophilus_influenzae	Host	Obligate	Extracellular	Pathogen	Non-motile
Helicobacter_pylori	Host	Obligate	Extracellular	Pathogen	Motile
Klebsiella_aerogenes	Mostly_host	Facultative	Extracellular	Both	Motile
Klebsiella_oxytoca	Mostly_host	Facultative	Extracellular	Both	Non-motile

Table S56: Information for key lifestyle traits for all 126 species in our dataset.
(continued)

Species	Host or free	Host association	Host location	Effect on host	Motility
Klebsiella_pneumoniae	Mostly_free	Facultative	Extracellular	Both	Non-motile
Lactobacillus_brevis	Mostly_free	Facultative	Extracellular	Mutualist	Non-motile
Lactobacillus_casei	Mostly_host	Facultative	Extracellular	Mutualist	Non-motile
Lactobacillus_delbrueckii	Mostly_host	Facultative	Extracellular	Mutualist	Non-motile
Lactobacillus_fermentum	Mostly_host	Facultative	Extracellular	Mutualist	Non-motile
Lactobacillus_helveticus	Mostly_host	Facultative	Extracellular	Mutualist	Non-motile
Lactobacillus_paracasei	Mostly_host	Facultative	Extracellular	Mutualist	Non-motile
Lactobacillus_plantarum	Mostly_host	Facultative	Extracellular	Mutualist	Non-motile
Lactobacillus_rhamnosus	Mostly_host	Facultative	Extracellular	Mutualist	Non-motile
Lactobacillus_sakei	Mostly_host	Facultative	Extracellular	Mutualist	Non-motile
Lactococcus_lactis	Mostly_host	Facultative	Extracellular	Both	Non-motile
Legionella_pneumophila	Mostly_host	Facultative	Intracellular	Pathogen	Motile
Leuconostoc_mesenteroides	Mostly_host	Facultative	Extracellular	Both	Non-motile
Listeria_monocytogenes	Mostly_free	Facultative	Both	Both	Motile
Mannheimia_haemolytica	Host	Obligate	Both	Both	Non-motile
Mycobacterium_avium	Mostly_host	Facultative	Both	Pathogen	Non-motile
Mycobacterium_bovis	Host	Obligate	Both	Pathogen	Non-motile
Mycobacterium_tuberculosis	Host	Obligate	Both	Pathogen	Non-motile
Mycobacteroides_abscessus	Mostly_free	Facultative	Both	Pathogen	Non-motile
Mycoplasma_bovis	Host	Obligate	Both	Pathogen	Non-motile
Mycoplasma_gallisepticum	Host	Obligate	Both	Pathogen	Motile
Mycoplasma_mycoides	Host	Obligate	Both	Pathogen	Non-motile
Mycoplasma_pneumoniae	Host	Obligate	Both	Pathogen	Motile
Neisseria_gonorrhoeae	Host	Obligate	Both	Pathogen	Non-motile
Neisseria_meningitidis	Host	Obligate	Both	Both	Non-motile
Paenibacillus_polymyxa	Mostly_host	Facultative	Both	Mutualist	Motile
Pasteurella_multocida	Host	Facultative	Extracellular	Both	Motile
Phaeobacter_inhibens	Mostly_free	Facultative	Both	Both	Motile
Piscirickettsia_salmonis	Host	Facultative	Intracellular	Pathogen	Non-motile
Porphyromonas_gingivalis	Host	Unknown	Both	Both	Non-motile
Prevotella_intermedia	Host	Unknown	Both	Both	Non-motile
Prochlorococcus_marinus	Free				Non-motile
Propionibacterium_freudenreichii	Mostly_free	Facultative	Extracellular	Mutualist	Non-motile
Proteus_mirabilis	Mostly_host	Facultative	Extracellular	Both	Motile
Pseudomonas_aeruginosa	Mostly_free	Facultative	Extracellular	Both	Motile
Pseudomonas_fluorescens	Mostly_host	Facultative	Extracellular	Both	Motile
Pseudomonas_putida	Mostly_free	Facultative	Extracellular	Both	Motile
Pseudomonas_stutzeri	Mostly_free	Facultative	Extracellular	Both	Motile
Pseudomonas_syringae	Host	Facultative	Extracellular	Pathogen	Motile
Ralstonia_solanacearum	Mostly_host	Facultative	Extracellular	Pathogen	Non-motile
Rhizobium_leguminosarum	Mostly_host	Facultative	Intracellular	Mutualist	Motile
Rhizobium_phaseoli	Mostly_host	Facultative	Intracellular	Mutualist	Motile
Rickettsia_japonica	Host	Obligate	Intracellular	Pathogen	Non-motile
Rickettsia_prowazekii	Host	Obligate	Intracellular	Pathogen	Non-motile
Rickettsia_rickettsii	Host	Obligate	Intracellular	Pathogen	Non-motile
Riemerella_anatipestifer	Host	Unknown	Both	Pathogen	Non-motile
Salmonella_enterica	Host	Facultative	Both	Pathogen	Motile
Serratia_marcescens	Mostly_free	Facultative	Intracellular	Pathogen	Motile
Sinorhizobium_meliloti	Mostly_host	Facultative	Intracellular	Mutualist	Motile
Staphylococcus_aureus	Host	Facultative	Extracellular	Both	Non-motile
Staphylococcus_epidermidis	Host	Facultative	Extracellular	Both	Non-motile
Stenotrophomonas_maltophilia	Mostly_free	Facultative	Extracellular	Both	Motile
Streptococcus_agalactiae	Host	Facultative	Extracellular	Both	Non-motile
Streptococcus_pneumoniae	Host	Facultative	Extracellular	Both	Non-motile
Streptococcus_pyogenes	Host	Facultative	Extracellular	Both	Non-motile
Streptococcus_suis	Host	Facultative	Extracellular	Both	Non-motile
Streptococcus_thermophilus	Unknown	Unknown			Non-motile
Vibrio_alginolyticus	Mostly_free	Facultative	Extracellular	Both	Motile
Vibrio_anguillarum	Mostly_free	Facultative	Extracellular	Both	Motile
Vibrio_cholerae	Mostly_free	Facultative	Extracellular	Pathogen	Motile

Table S56: Information for key lifestyle traits for all 126 species in our dataset. *(continued)*

Species	Host or free	Host association	Host location	Effect on host	Motility
Vibrio_parahaemolyticus	Mostly_free	Facultative	Extracellular	Both	Motile
Vibrio_vulnificus	Mostly_free	Facultative	Extracellular	Both	Motile
Xanthomonas_campestris	Host	Facultative	Extracellular	Pathogen	Motile
Xanthomonas_citri	Host	Facultative	Extracellular	Pathogen	Motile
Xanthomonas_oryzae	Host	Facultative	Extracellular	Pathogen	Motile
Xylella_fastidiosa	Host	Facultative	Extracellular	Pathogen	Motile
Yersinia_enterocolitica	Mostly_host	Facultative	Both	Pathogen	Both
Yersinia_pestis	Mostly_host	Facultative	Both	Pathogen	Non-motile
Yersinia_pseudotuberculosis	Mostly_host	Facultative	Both	Pathogen	Both
Zymomonas_mobilis	Mostly_host	Unknown	Unknown	Unknown	Both