Universal bacterial clade dynamics dominate under predation despite altered phenotypes and mutation targets

Dovydas Kičiatovas[1], Johannes Cairns[1,2,3], Veera Partanen[4], Julius Hoffmann[5], Lutz Becks[5,*], Teppo Hiltunen[2,*], and Ville Mustonen[1,*]

[1]Department of Computer Science, Organismal and Evolutionary Biology Research Programme, University of Helsinki, Helsinki, Finland
[2]Department of Biology, University of Turku, Turku, Finland
[3]Turku Collegium for Science, Medicine and Technology, Turku, Finland
[4]Department of Microbiology, University of Helsinki, Helsinki, Finland
[5]Department of Biology, University of Konstanz, Konstanz, Germany

February 4, 2025

Abstract

Recent studies have revealed bacterial genome-wide evolution to be complex and dynamic even in a constant environment. The evolution is characterized by the emergence of new clades competing or temporarily coexisting as each clade continues to undergo evolutionary change. Previous studies on predator-prey dynamics tracking simple ecological and phenotypic metrics have shown predation to fundamentally alter prey evolution, often giving rise to defense evolution followed by frequency dependent selection between defended and undefended prey genotypes as well as coevolution. Moreover, these dynamics can differ between prey species depending, for example, on mutation rate and the physiological and fitness effects of the mutations available. Here we sought to consolidate these fields by examining genome-wide evolution in five bacterial prey species separately subjected to long-term evolution under ciliate predation. For all species, we found mutational signals of prey defense evolution, with phenotypic data demonstrating change in phenotypic composition for defense trait over time. Mutational targets and dynamics also differed between prey species. Intriguingly, however, overall temporal molecular dynamics across the prey species were strikingly similar to those of bacteria evolving alone, with constant emergence, competition and quasi-stable coexistence of clades. This makes molecular evolution more interesting and less predictable than we might expect based on existing coevolutionary theories.

Main

Understanding how organisms adapt and evolve is a fundamental challenge in biology that continues to intrigue researchers. While organisms seemingly respond to environmental pressures through predictable phenotypic changes[1], the underlying genetic mechanisms reveal a far more complex and unpredictable reality[2]. At the phenotypic level, adaptation is often repeatable and predictable, both in controlled experiments[3,4] and when comparing across natural populations[5]. This repeatability can allow us to identify key mechanisms driving adaptive changes, such as selection strength[6,7] and trade-offs between traits[8,9]. However, at the nucleotide level, evolution is typically much less predictable, even under strong selective pressures[10]. This unpredictability arises from two main factors. First, different metabolic pathways can lead to the same phenotypic outcome, allowing for multiple routes of adaptation[11,12]. Second, even in relatively simple environments and within a single, isolated population, clonal structure can undergo continuous and significant changes over time. These changes are driven by complex dynamics, such as competition and persistence among mutational or clonal cohorts, which can coexist for extended periods[13,14]. Clonal cohorts are groups of genetically identical individuals that arise from a common ancestor and share the same mutations. These cohorts play a crucial role in shaping the genetic composition of populations over time, as they compete, persist, or decline based on their fitness advantages or disadvantages, and chance. Thus, molecular evolution is shaped not only by direct selection and the emergence of new mutations but also by the complex interactions among clonal cohorts within populations over generations. Understanding these dynamics is crucial for uncovering how genetic changes respond to environmental pressures and why evolutionary outcomes can diverge, even under similar selective conditions.

Ecological interactions, those between individuals of different species, introduce additional layers of complexity to molecular evolution[15]. For instance, species interactions can drive strong directional or stabilizing selection on populations. Strong selection may increase the efficiency of eliminating competing clonal lineages, simplifying clonal population dynamics[16]. On the other hand, species interactions can lead to coevolution and eco-evolutionary dynamics, where changes in population size both drive and result from evolutionary shifts occurring on similar timescales[17]. In these scenarios, selection pressures fluctuate continuously, varying in direction and/or strength. These fluctuations can promote the coexistence of clonal cohorts in a quasi-stable or frequency-dependent manner, making molecular evolution at least as dynamic as in single-species systems. For example, a study of microbial host-virus coevolution found that the host harbored significant genetic diversity, primarily driven by neutral molecular evolution rather than direct selection by the virus. The virus underwent fewer, strongly selected mutations, typically confined to specific genes[18]. Furthermore, the complexity of species interactions can lead to conflicting selection pressures, potentially altering the trajectory of molecular evolution. This has been observed in tripartite interactions involving bacteria, phages, and plasmids, where the presence of multiple mobile genetic elements constrains the evolutionary responses typically seen in pairwise interactions[19]. Coevolution can also accelerate molecular evolution[20], driving successive sweeps of beneficial mutations and, in some cases, leading to simpler evolutionary dynamics. These examples illustrates how coevolutionary dynamics can lead to both complex or more simple patterns of molecular evolution. Since all populations are part of larger communities and interact with other species, these interactions play a fundamental role in shaping evolutionary dynamics. Therefore, it is crucial to study their influence on molecular evolution. Understanding these dynamics is essential for predicting and managing evolutionary outcomes in natural and artificial systems, from ecosystem conservation to the development of sustainable disease control strategies[21].

The patterns and rates of molecular evolution differ depending on the timescale observed. Experimental evolution studies, which track molecular evolution in replicated populations, have shown that, over short timescales, beneficial mutations accumulate rapidly in response to new selective pressures. These mutations are often large-effect and lead to significant fitness gains in a short period, allowing for an evaluation of predictability in evolutionary processes[22]. Long-term dynamics, however, are more likely governed by declining rates of adaptation, as beneficial mutations become rarer, while the overall rate of molecular evolution remains constant. Although much of our understanding of molecular evolution in the context of species interactions comes from short-term experiments, the more complex evolutionary dynamics are only observable over longer periods. Unfortunately, such long-term studies are currently lacking, leaving unanswered questions about how species interactions influence molecular evolution.

In this study, we investigated the long-term molecular and phenotypic evolution of five bacterial species (*Brevundimonas diminuta*, *Comamonas testosteroni*, *Pseudomonas fluorescens*, *Sphingomonas capsulata*, and *Serratia marcescens*) as they grew with and without a ciliate predator, *Tetrahymena thermophila*, over a period of 800 days (Fig. 1; see Methods). Our aim was to quantitatively assess how the presence of the predator influenced molecular evolution in the bacterial prey. Specifically, we sought to determine whether the predator's presence led to measurable changes in long-term molecular evolution compared to predator-free populations. To achieve this, we tracked clonal diversity and phenotypic changes in replicated bacterial populations over time, identifying recurrent gene targets along the way.

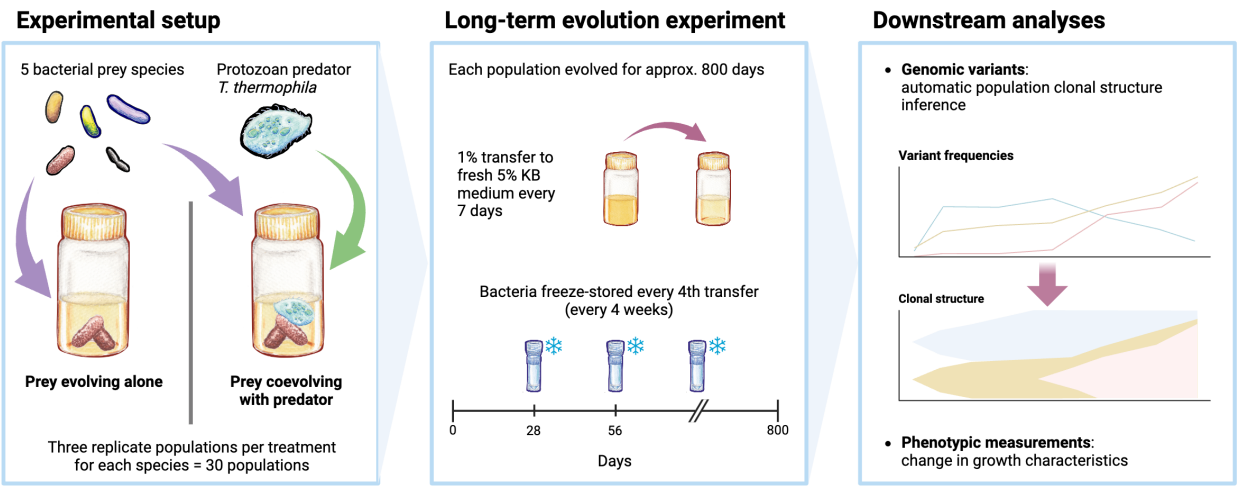

Figure 1: Graphical abstract.

Results

Growth phenotypes differ between prey populations coevolving with a predator and populations evolving alone

First, we investigated the effects of predation on the prey population growth dynamics. A previous study by Cairns *et al.* on the same predator-prey system[23] has shown that the ciliates evolve alongside the bacterial prey. Here, we sought to describe the phenotypic changes in these prey species over a longer period of time than previously studied. We analysed the growth curve data (i.e., growth phenotype) of sampled populations from each study population at the beginning and the end of the long-term experiment. The isolates were grown on three nutrient-rich experimental media – without ciliate (control), with ciliate and with added KCl salt without ciliate (see Methods). Adaptation to a particular stressor typically influences multiple bacterial phenotypes due to the pleiotropic effects of altering key structures or functions, as is the case with antibiotic cross resistance and collateral sensitivity[24]. Therefore, in addition to studying the phenotypes of bacteria evolved alone or with ciliate in the primary conditions of interest – experimental medium with and without ciliate predation – we also included a salt stress condition.

For each growth curve, we computed yield Y, maximal per-capita growth rate ρ_{max}, time to the latter $t_{\rho_{max}}$ and the area under the curve (AUC). We contrasted these parameters between the coevolutionary histories and found that all of their means differ significantly when the bacterium is grown in a medium where the ciliate is present. Yield, AUC and ρ_{max} are higher, and the latter is attained sooner when the bacteria have coevolved together with the predator over the course of the long-term evolutionary experiment, meaning that in the presence of predation, they consistently grow better than bacterium that evolved alone. This is expected as bacteria evolve anti-predatory defenses in the presence of the ciliate, as shown before in a previously described ciliate-*P. fluorescens* system in Scheuerl *et al.* study[25]. However, there is no sufficiently strong statistical evidence that growth differs in either control or salt stress media. Nevertheless, it is noteworthy that the coevolved bacteria isolates do not display weaker growth capabilities in the control medium, compared to the bacteria that evolved alone. This could be explained by the expectation that coevolved prey populations would include both more reproductionally fit undefended phenotypes and ciliate-resistant but less fit defended phenotypes.

We did not observe a consistent direction of growth phenotype evolution with respect to time in the long-term experiment. Isolates sampled later in the experiment may grow better or worse than in the beginning without a clear pattern across replicates (Supplementary Fig. 4). Despite the overall signal separating coevolutionary histories in terms of different maximal rate of per-capita growth in the presence of predation, each species has a unique phenotypic response to each growth medium, regardless of their coevolutionary history. Notably, all replicates of both treatments of *S. marcescens* do not appear to have any notable difference when grown in the presence of predation, possibly indicating pre-existing defenses to ciliate (Supplementary Fig. 4, bottom row). Isolates of *B. diminuta*, *P. fluorescens* and *S. capsulata* show the most stark difference in growth rates between the histories in the presence of predation. Overall, the variance between isolates' growth is lower in the presence of predation than in the other test conditions. To summarize the differences in growth between the coevolutionary histories in terms of maximal per-capita growth rate and total yield, we computed two ratios between isolates grown in the control medium and the experimental medium with predator: ratio of yields and ratio of maximal per-capita growth rates (Fig. 2B). These ratios can be in-

terpreted as growth advantage/disadvantage of isolates grown in the control medium as opposed to growth in the experimental medium with predator (advantage in this medium is represented through lower values of either ratio). Three species – *B. diminuta*, *C. testosteroni* and *P. fluorescens* – have different variances and means in both ratios (two-sided Kolmogorov-Smirnov test: *B. diminuta* yield ratio p-value < 0.001, ρ_{max} ratio p-value < 0.001; *C. testosteroni* yield ratio p-value < 0.001, ρ_{max} ratio p-value < 0.001; *P. fluorescens* yield ratio p-value < 0.001, ρ_{max} ratio p-value < 0.001), further suggesting more predictable growth when prey populations are coevolved. However, the evidence for the difference in ρ_{max} ratios in *S. capsulata* and *S. marcescens* populations is weaker (two-sided Kolmogorov-Smirnov test: *S. capsulata* yield ratio p-value < 0.001, ρ_{max} ratio p-value $= 0.12$; *S. marcescens* yield ratio p-value < 0.001, ρ_{max} ratio p-value $= 0.002$). Overall, it appears that isolates from coevolved populations are more homogeneous in terms of their growth advantage.

To assess the predictability of the growth curves in terms of the species, coevolutionary history, replicate population within species and the timing of the isolate sampling, we constructed a label for each growth curve, comprised of these categorical variables. Using the growth curves, concatenated across the three types of media the isolates were grown in as our input, we trained an ensemble of 100 random forest models with random splits of 50% of the data, predicted the label of the remaining randomly split 50% of growth curves and averaged the prediction results across the random forest instances (see Methods; Fig. 3).

As expected, bacterial adaptation in the presence or absence of ciliate predation had a multidimensional effect on bacterial phenotypes. Moreover, this effect was specific to the bacterial species, and within species, to the coevolutionary history. However, the replicate populations are not as distinct within the pooled isolates of the same coevolutionary history, and neither across the time points within a replicate population. Nevertheless, we clearly see higher classification accuracy between the three coevolved populations in every species (Fig. 3, right hand panel) than in populations that evolved without the predator, suggesting that the phenotypic dynamics are more distinguishable between coevolved populations.

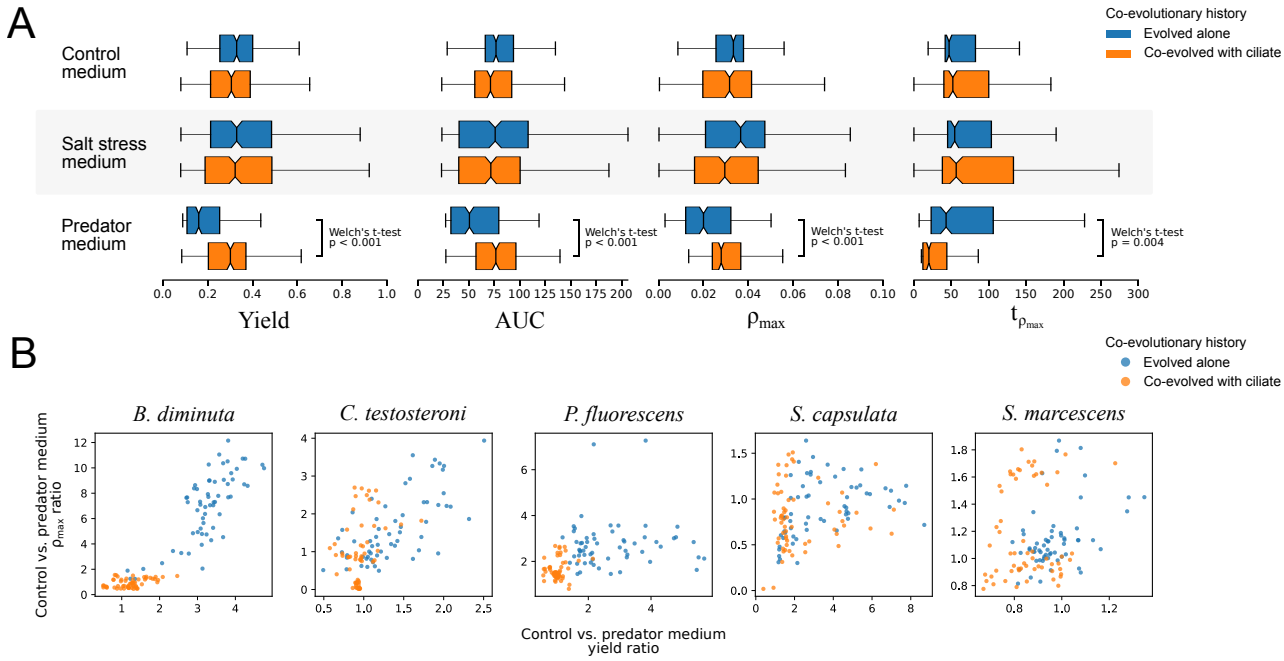

Figure 2: **A** Growth curve statistics: yield (optical density at the end of measurement), area under the curve (AUC), maximal per-capita growth rate ρ_{max} and time to maximal per-capita growth rate $t_{\rho_{max}}$. Statistically significant results from two-sided Welch's t-tests, comparing the means of all isolates between opposite coevolutionary histories when grown in the three different media (see Methods), are marked next to the relevant pair of boxes. **B** A scatter-plot of the ratios of two growth phenotype metrics – yield and maximal per-capita growth rate ρ_{max} – between the control medium and the experimental medium with the predator.

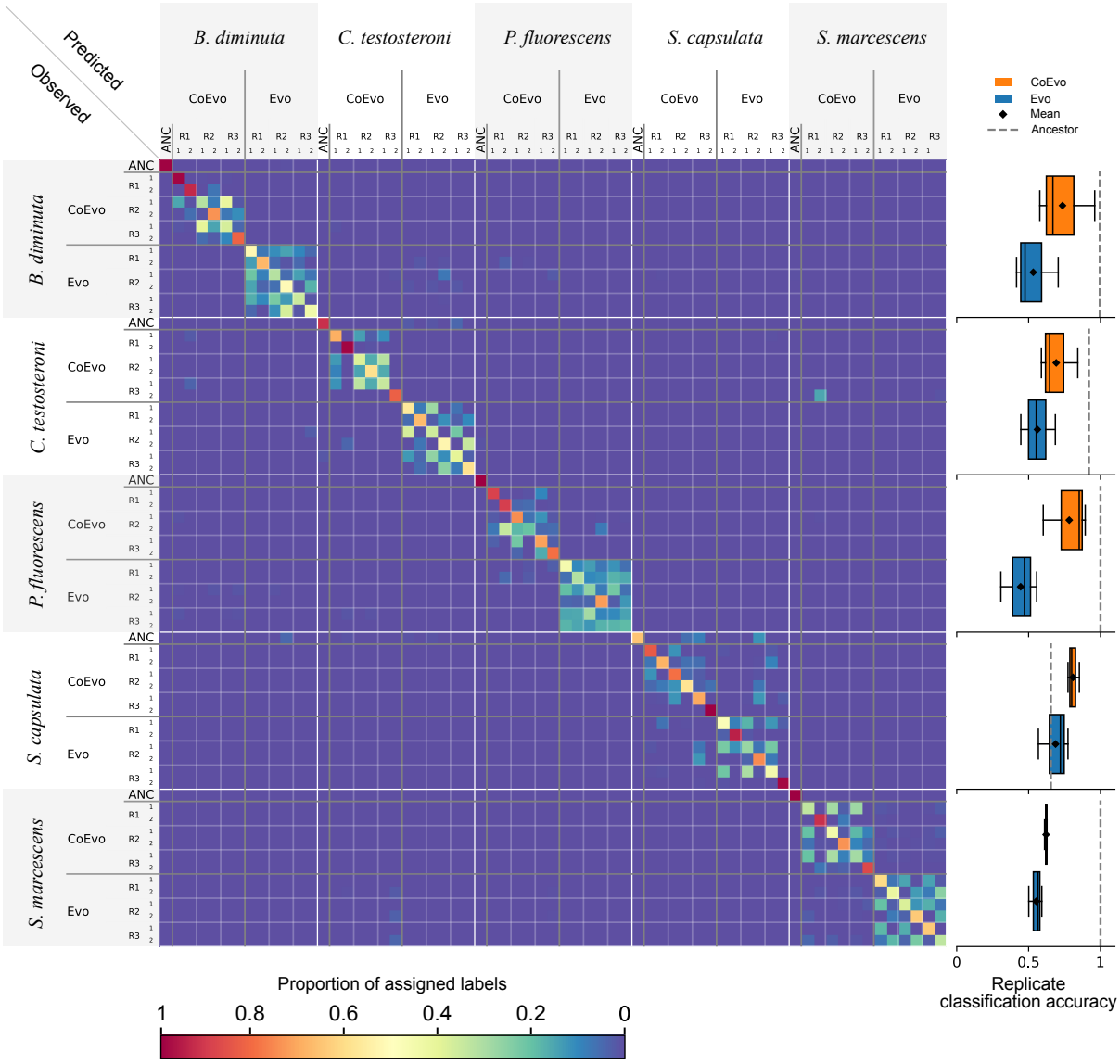

Figure 3: Combined and averaged confusion matrix of 100 random forest classifier instances, trained on 50% of randomly sampled isolate growth profiles (combined isolate growth curves in all three growth media) and tested on the remaining 50% of the data (label proportions preserved in both training and testing sets). The X-axis represents actual data labels and the Y-axis – predicted. The rows and columns of the confusion matrix are separated according to the alphabetic order of the species names, and each inner square of the confusion matrix, representing a species, is further subdivided into the two evolutionary histories, then three replicate populations and lastly, the time points. The color gradient represents the average number of assigned samples from the test set, averaged over 100 random forest instances. While there is little confusion between the species and evolutionary histories, the growth profiles of replicate populations within species and their evolutionary history are more difficult to separate. However, coevolved species-specific replicates are distinguished with higher accuracy than when evolved alone (right panel).

Bacterial molecular evolutionary dynamics are similar regardless of the presence or absence of predator

Next, for each replicate of a given prey species, evolving with and without the predator present, we calculated the frequencies of single nucleotide polymorphisms (SNPs) and short insertions or deletions (indels) and their change over the course of the experiment. The final accumulated variant counts differ vastly between species and between replicate populations (Supplementary Fig. 1, Supplementary Table 1), indicating a large degree of variability in the accumulated mutation counts even within the replicates of the same species under same environmental conditions. Likewise, population 1 of *C. testosteroni* steadily acquires almost double the number of variants compared to the remaining populations of either treatment of the same species. However, at the 95% confidence level, none of the species has a significant two-sided Welch's t-test result when comparing the means of final accumulated variants counts between experimental groups – although it is worth noting that the sample sizes for comparison might be too low. The trajectories of most variants form clearly defined clusters, in line with the expectation that most mutations would hitchhike with beneficial mutations under selection[26]. Henceforth, we

analyzed these clusters as variant cohorts (see Methods; Fig. 4, main panel).

Several moderate to high-frequency variants in known mutator genes were observed together with a sharp increase in mutation counts right after the time of their detection, likely explaining why some study populations accrue much higher mutation counts than others within the same species and treatment. Variants in mutator gene *mutS* were detected in three species in both treatments: a stop-gain (appeared on day 700, mean frequency after appearance 0.32) and disruptive inframe deletion (appeared on day 476, mean frequency after appearance 0.36) in coevolved *P. fluorescens* populations 2 and 3, respectively; a missense variant in coevolved *S. marcescens* population 2 (appeared on day 679, mean frequency after appearance 0.75); disruptive (appeared on day 553, mean frequency after appearance 0.99) and conservative (appeared on day 553, mean frequency after appearance 0.96) inframe deletions in evolved-alone *C. testosteroni* populations 2 and 3, respectively; and a missense variant in evolved-alone *P. fluorescens* population 1 (appeared on day 412, mean frequency after appearance 0.64). Variants in *mutL* were detected in the following populations: a frameshift variant in coevolved *C. testosteroni* population 1 (appeared on day 448, mean frequency after appearance 0.35); a frameshift variant in evolved-alone *C. testosteroni* population 1 (appeared on day 146, mean frequency after appearance 0.66); and a stop-gain variant in evolved-alone *B. diminuta* population 1 (appeared on day 272, mean frequency after appearance 0.65). Additionally, in *C. testosteroni*, coevolved population 1 had a *uvrC* missense variant (appeared on day 504, mean frequency after appearance 0.13) and evolved-alone population 1 had two *uvrA* missense variants (appeared on days 553 and 776, mean frequencies after appearance 0.55 and 0.25).

Next, using the variant cohorts, we inferred the clonal structure within each population at every time point – treating the cohorts as genetic alleles, we developed a tree-based algorithm to assign each cohort to the role of either a new clone or a sub-clone of an existing clone (see Methods). Reconstructing clonal structures from variant frequency data is widely used in cancer research, where linking cancer evolution to clinical outcomes can inform therapies[27,28]. Cancer clonal structure inference is difficult due to several factors, such as DNA copy number variation, genome diploidy and paucity of long term time-series data[29], which are not present in our bacterial genomes, making this type of inference more readily viable. With the inferred allele hierarchy, we calculated the allele frequencies over time and visualize them as Muller plots and phylogenetic trees (Fig. 4). Muller plots in this kind of population structure inference from variant data have been presented before[30], however, to our knowledge, there is no tool that computes high-confidence population clonal structures automatically (nevertheless, we have manually checked all our visualised solutions as a form of quality control, see Methods).

We observed a large degree of variability in clonal dynamics, even between the replicates of the same species (Fig. 4, Muller plots). Most prey species exhibit similar variant frequency dynamics, such as selective sweeps that occur throughout the entire duration of the experiment. We have also observed non-trivial dynamics of variant clusters being maintained at a high frequency over long periods of time without these mutations getting lost or fixed (Fig. 4, Muller plots), suggesting previously described state of quasi-stable coexistence in bacterial populations[13]. The latter is clearly visible in the population 1 of *B. diminuta* in clades AC and AD, both undergoing their own sub-clonal expansions at different points in time. Other species exhibit similar clade coexistence, albeit with more stark difference in the clade frequencies. Similarly, selective sweeps occur in several study replicates, with coevolved *S. capsulata* population 3 displaying textbook successive sweeps. Successive hard selective sweeps are only observed in four cases; interestingly, all three coevolved *S. capsulata* populations, grown with ciliate, are among them, and the remaining one is a replicate of *C. testosteroni* grown without ciliate (Fig. 4, "Coevolved with ciliate" and "Evolved alone" panels). Furthermore, many populations show a pattern of late large scale diversification events, where many different clades start to appear in a quick succession halfway into the experiment (more than 400 days).

The number of clades consistently in both treatments increases over the duration of the experiment to the maximum of nine clades (linear regression coefficient of determination $R^2 = 0.75$ and slope 0.0085 in coevolved populations, $R^2 = 0.69$ and slope 0.0081 in evolved-alone populations). However, interestingly, the number of clades present at any given time in the experiment, after approximately 250 days, consistently stays at 3-5 clades (Supplementary Fig. 2) – suggesting active turnover of clades within a population as time goes on. This is likely explained by clonal interference between clades – however, since we do not observe consistent and repeated selective sweeps across the whole population, clonal interference dynamics are likely moderated by frequency dependent selection dynamics that sustain clade coexistence for longer periods of time[13,31].

To quantitatively compare the inferred clonal structures between the experimental groups, in addition to Muller plots, we computed hierarchical representations in the form of strictly bifurcating phylogenetic trees (Fig. 4, main panel, Supplementary Fig. 6, Supplementary Fig. 7), where the total branch length represents the number of days a clone existed since its emergence until either the end of the studied experiment or the final point in time the clone retains a frequency above 0.1 (provided no further sub-clones appear) – otherwise, the branch length is truncated.

We computed the sum, mean and variance of branch lengths, the number of leaves (corresponds to number of clones plus the background clone), maximum and average depths of the leaves, as well as Sackin, Colless and B1 indices of tree balance[32]. Using regression analysis, we tested these metrics for statistically significant differences between the experimental groups (with predator and without predator) in two ways: across all study replicates, using the species indicator as a covariate (ANCOVA[33]), and between species-specific experimental groups (ANOVA) – interestingly, none of these measures differ significantly between the treatments (95% CL).

Going by species, only the experimental groups of *B. diminuta* are significantly differing in their trees' B1 indices (at 95%CL, p-value 0.004) – trees that have their leaves closer to the root on average are considered balanced, indicating that there is a greater imbalance observed in the structure of populations that evolved without ciliate. For complete table of within-species ANOVA results, see Supplementary Table 2.

These results suggest that the presence of ciliate does not predictably affect the clonal structure of the studied prey populations, or, at the very least, the changes are subtle and remain largely hidden under shared evolutionary processes, such as adaptation to the growth medium. Instead, we observed the usual clonal interference and likely frequency dependent selection dynamics that universally characterize even simple microbial systems without predation[31].

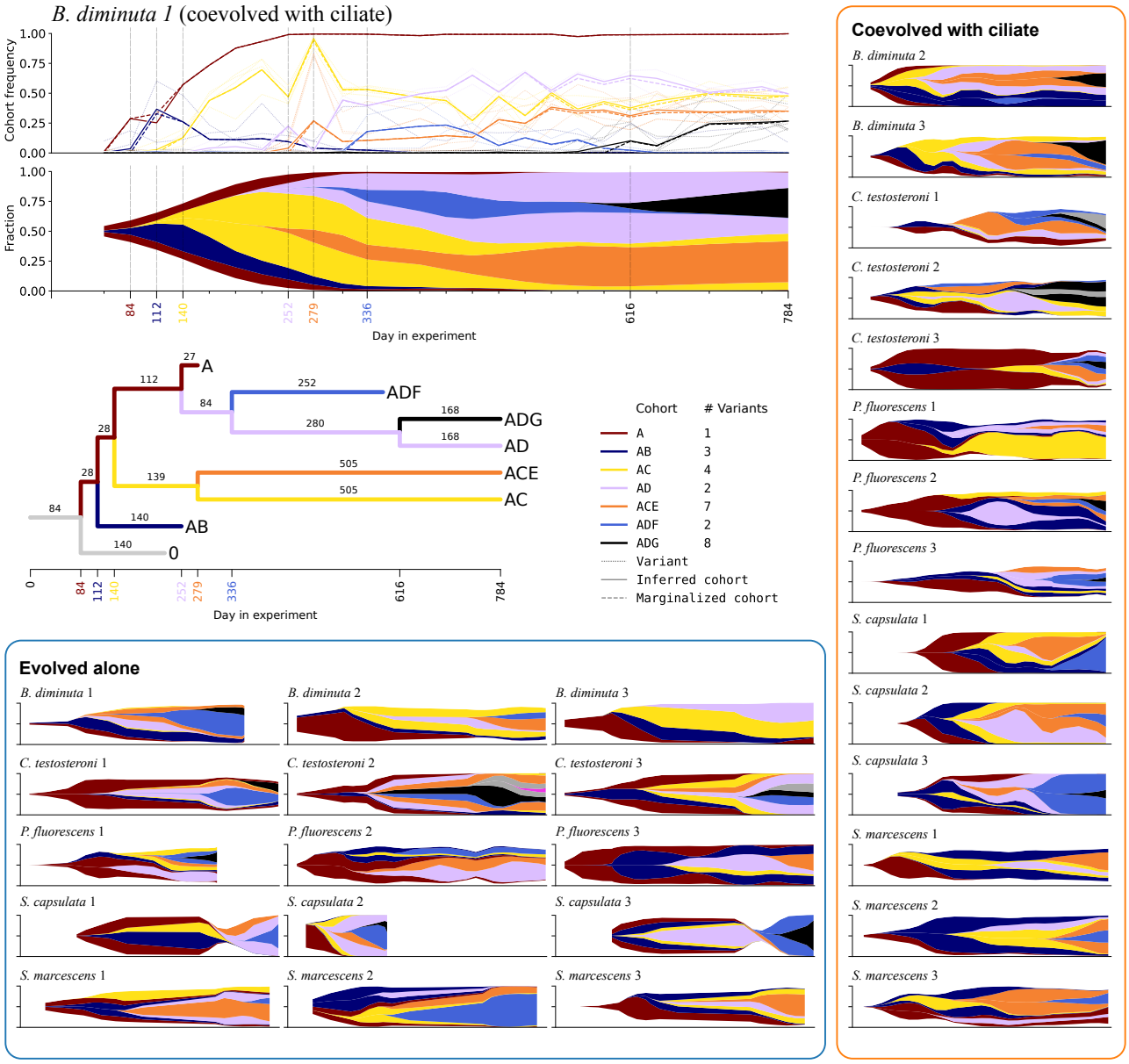

Figure 4: Clonal structure evolution of bacterial populations under and without predation in terms of variant frequency trajectories over the course of the long-term experiment. Trajectories are first clustered into cohorts and then assigned to a likely position in the clonal structure (see Methods) and represented as both Muller plots and phylogenetic trees. An example is given for clonal dynamics of *B. diminuta* replicate population 1. Vertical dashed bars in the trajectory and Muller plots (upper left, first two panels, respectively) mark the appearance of the clone (cohort trajectory exceeding the emergence threshold, see Methods). The time of the appearance of each clone is color-coded to correspond to the clone information in the legend and also the phylogenetic tree. The branch labels in the tree denote the time the clone remains above a survival threshold (provided no further sub-clones emerge; see Methods). A non-trivial long-term coexistence of the inferred clades is clearly seen in the Muller plot and the phylogenetic tree. Orange and blue boxes: Muller plot showcase for the remaining 29 populations evolving under and without predation, respectively. There is significant variability both between species and between replicate populations of the same species. Long-term coexistence of clades is a recurring theme, along with examples of selective sweeps. Populations with truncated Muller plots represent populations where a portion of time series data was lost to contamination during laboratory procedures.

Mutation targets are recurrent both within and across species, and differ depending on the presence or absence of the predator

Lastly, we investigated the recurrence of non-synonymous variants identified by our genomic analyses to determine if there are patterns in recurrence and evidence of selection for mutations in specific genes.

We define the recurrence of a mutated gene as it being affected by at least one variant in at least two out of three replicates of a species, with and without the presence of the predator separately (Fig. 5, top panel).

We also investigated the appearance times of the recurrent gene mutations, i.e., the appearance of the clade they belong in (Fig. 5, bottom panel) and found that the overall distribution of the timing of appearance is significantly differing between the coevolutionary histories, i.e., populations evolving with the

predator and populations evolving alone (two-sided Kolmogorov-Smirnov test p-value = 0.00003; Fig. 5, bottom right histogram). It appears that in the populations coevolving with the predator, more recurrently targeted genes are affected late in the experiment (> 600 days) than in populations evolving alone, moreover, only the coevolving populations show recurrently affected genes appearing within the first 100 days of the experiment.

In total, only eight genes – *ppk*, *sasA*, *glpK*, *rcsC*, *mutS* *dppA*, *rpoB* and *corA* – are recurrent in more than one of the species. The gene encoding the adaptive sensory kinase *sasA*, recurrently affected in all studied species except *S. marcescens*, is purported to be a core part of the circadian rhythm control mechanism in cyanobacteria[34], though to our knowledge, its exact physiological role in other bacteria is not known. Polyphosphate kinase gene *ppk* is responsible for a variety of functions[35] and is recurrently targeted in 3 species, though the recurrence is not limited to coevolved populations, unlike *sasA*. All remaining affected genes are limited to recurrence between at most two species.

To test whether the recurrence of any gene is statistically significant, we performed a regression analysis using ANCOVA on gene hit indicator data (1 if a non-synonymous variant has been detected in a gene at least once, 0 otherwise) between with- and no-predator groups using species indicator as a covariate. Nine genes – *barA*, *dppB*, *flhC*, *lptA*, *rcsC*, *relA*, *rpoB*, *rpoC* and *sasA* – have been identified as having a significantly differing recurrence between the experimental groups, i.e., growth with ciliate versus without ciliate (Supplementary Table 3).

The majority of remaining genes recurrent only in coevolved populations are responsible for various metabolic functions, cell wall formation and maintenance, though some are related to biofilm formation, which is expected given bacteria propensity to form biofilm in response to external stress, such as predation, in addition to simply being a viable strategy to adapt to the environment[36]. Genes *flhC* and *rpoC*, recurring only in coevolved *S. marcescens* populations, are associated with biofilm formation[37–39]. Hybrid sensor kinase gene *rcsC* has been identified as an important part of *E. coli* biofilm formation[37] and is recurrently affected in all replicates of both *C. testosteroni* and *P. fluorescens*. In contrast, genes *rpoB*, *pilQ* and *relA* are associated with biofilms and motility[39,40] and are recurrent only in *C. testosteroni* populations grown without ciliate and are likely not related to defense against the predator. Gene *rpoB* is also recurrent in two out of three *B. diminuta* evolved-alone populations. Gene *cheB* (also known as *wspF*) is recurrent in all three coevolved *P. fluorescens* populations and has been previously reported to be also recurrently targeted in a dual-stressor (antibiotic and ciliate) system with *P. fluorescens* as prey[41]. Additionally, mutator gene *mutS* is recurrent in *P. fluorescens* populations. Many genes that recur in both experimental treatments are related to various metabolic functions and are likely responsible for adaptation to the growth medium rather than defense.

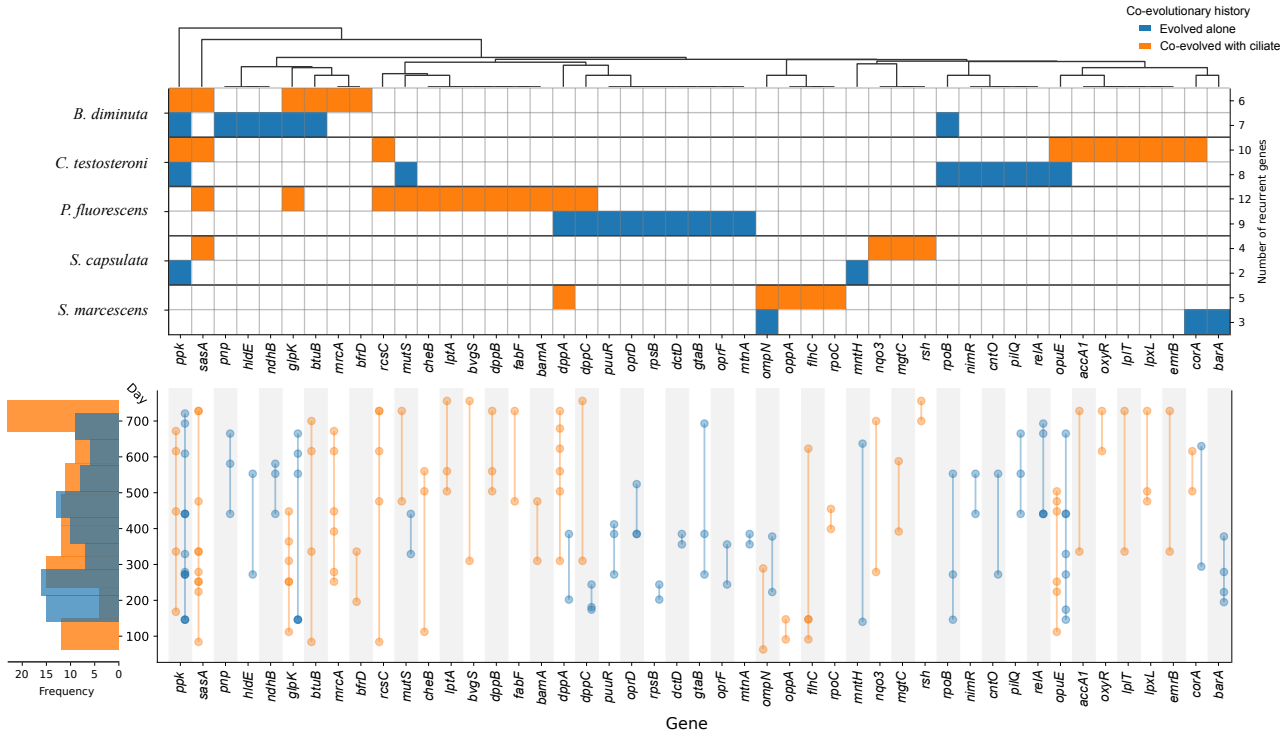

Figure 5: Top panel: non-synonymous variant hit recurrency across the experiments in known genes. Columns (i.e., genes) are hierarchically clustered according to the pairwise Hamming distance. Populations without ciliate (blue rows) have 17 recurrent genes containing variants that do not occur in populations grown with the ciliate (orange rows). Conversely, the set of genes that are hit only in populations with ciliate is larger. Bottom panel: variant timing of appearance (based on cohort appearance to which a variant belongs to). Darker circles indicate that there are multiple unique study replicates where a variant in a given gene appeared at the same time. Marginal counts of variant appearance timing are shown as a histogram.

Discussion

Our study builds upon the long-term experimental evolution methodology presented in Good *et al.* study[13], where bacterial molecular evolution is shown to be complex and incessant over long time scales, even in a constant environment. With our experiment, we aimed to study whether a strong selective pressure, i.e., predation, would significantly change the clonal dynamics and the outcome of molecular adaptation in the prey bacterium. We observed clear signs of phenotypic adaptation, where coevolved bacterial prey populations grow significantly better in the presence of ciliate compared to control populations. Moreover, recurrently mutated gene sets are different between the experimental groups – both phenotypic and genomic data support treatment-specific evolution. Population sizes of predator and prey in coevolved populations, as expected, fluctuate in opposition to one another and are negatively correlated (linear regression $R^2 = -0.42$, slope -1.78, Supplementary Fig. 5).

However, surprisingly, despite our observance of rich and non-trivial clonal structure dynamics, we cannot detect a strong signal of predation within the clonal structures of our study populations. Instead, we find that the molecular evolutionary dynamics are predominated by processes common to all environments. These processes feature constant dynamic emergence of adaptive clades, and coexistence and competitive dynamics between them[13].

What we observed in this study may be a case of "emergent simplicity" – complex and unpredictable genomic evolution is ultimately resulting in predictable phenotypic changes[10]. There may be many metabolic pathways through which prey bacterium adapt to their immediate environment (both the growth medium and the ciliate), changing their mode of growth in a predictable manner, but on a genomic level, prey population clonal dynamics may be largely driven by the aforementioned universal evolutionary processes, much like in previous studies on pairwise bacterial community evolution[42]. Partly predictable evolution in terms of clade dynamics has been described in influenza[43,44], where strong clonal interference drives successful clades to constantly emerge victorious over the other clades in the population – that is predominantly not the case in our clade trees, and instead we observed constant competition and coexistence of clades. An interesting thing to note is that we observe rapid changes in clonal diversity that occur rather late – in some cases, even in the final 100 days in our 800 day-long experiment (Fig. 5) – so it would be interesting to see whether there are any significant changes in clonal structure over much longer timescales, though that can be, unfortunately, time and cost prohibitive.

There might be several explanations for the lack of clearly observable predator signal in clade dynamics in our study: 1) our data resolution might be insuf-

ficient to observe subtle changes in between genomic sampling (every four weeks in our study); 2) the predator in our study, *Tetrahymena Thermophila*, is a generalist predator, and therefore may not exert a strong enough selective pressure to influence clade dynamics; and 3) frequency dependent selection, which prevents a quick competitive exclusion of the prey clades[31], unlike in influenza evolution. Frequency dependent selection is further supported by our observance of higher phenotypic heterogeneity in isolates that have evolved without a predator, when grown in the experimental medium with the predator present (Fig. 2B). Moreover, within each population, there is a constant maintenance of several clades coexisting at the same time above threshold frequencies (Supplementary Fig. 2).

Also, we recognize that our statistical power (three replicates per species, per treatment) might be a limiting factor in detecting subtle clonal dynamics. To fully disentangle the clonal dynamics of coevolving prey, we believe that a larger number of phenotypic measurements (than 10 clones per sampling time point in our study) would be needed from every sampling time point to account for the ever-increasing clonal diversity of the populations – growth profiles of the limited number of isolates taken from a time point where one clone dominates might be hiding subtle effects of clones that are present at lower frequencies. Additionally, since the first time point of isolate sampling represents populations that have already (co-) evolved for at least a month, we cannot compare early prey growth dynamics between the treatments. Furthermore, sampling the populations for genomic data every four weeks might be too coarse to account for rapid eco-evolutionary dynamics, such as overall prey and predator population size fluctuations during one-week transfer cycles (see Methods), though it is not clear whether increasing the overall time resolution with respect to the genomic data sampling alone would result in different conclusions.

For future studies, we believe that long-term coevolution experiments need carefully designed linking between genomic and phenotypic data tracks in order to disentangle the impact of strong environmental pressure (i.e., predation) from other evolutionary processes. Ignoring the identity of mutated genes, clonal dynamics, represented by variant cohort dynamics of SNPs and indels, seems quickly to become so complex that differentiating predation pressure from other evolutionary processes molding it remains an outstanding challenge. Furthermore, it would be interesting to study the coevolution of the predator as well and contrast the evolution of prey defense to evolution of predator offense – for that, one would need to produce equally rich genomic data from the predator and investigate clonal dynamics and gene recurrence in a similar fashion, though it is expected that the prey would experience greater evolutionary changes than the predator. Nevertheless, studying predator coevolution in terms of predator population clade dynamics might yield interesting novel insights.

While it is still surprising that in this study we do not observe differences in clade dynamics between our evolutionary treatments, given the strong phenotypic and genomic effects from predation, we believe that the predator signals, while being detectable, remain masked within the clade dynamics by pervasive prey clade interactions, stable coexistence and frequency dependent selection that maintains both undefended and defended types within the prey populations. We believe that our study further reinforces the non-triviality of clonal dynamics even in the presence of a strong selective pressure despite easily detectable phenotypic differences and also provides new methodology to represent and investigate the clonal structure of a population solely through genomic time series data.

Materials and Methods

Experimental setup

The focal predator of this experiment is a ciliate *Tetrahymena thermophila*, preying upon these five bacterial species: *Brevundimonas diminuta*, *Comamonas testosteroni*, *Pseudomonas fluorescens*, *Sphingomonas capsulata* and *Serratia marcescens*. *Brevundimonas diminuta* has been previously studied and is found in lake habitats shared with *Tetrahymena*[45], *Comamonas testosteroni* is known to have specific defenses against protozoan grazing[46]. The remaining three – *Pseudomonas fluorescens*, *Sphingomonas capsulata* and *Serratia marcescens* – are model species commonly used in prey–predator studies. They are a subset of the same organisms studied in the Cairns *et al.* 2020 study[23]; therefore, a detailed description of these species is omitted in this article.

For each prey species, three replicate populations were prepared and grown in 5% King's Broth (KB) medium with and without the presence of the ciliate. In total, 30 bacterial populations have been evolved over the course of 800 days, with 1% of the population being transferred to a fresh medium every 7 days. A sample from each prey population was taken and freeze-stored every 4 transfers (i.e., 28 days), and predator cells were made axenic and stored in liquid nitrogen 1–2 times per year. Optical densities (prey) and cell counts (predator) were recorded at every transfer to obtain a time series of population growth over the course of the experiment.

Genomic data and variant calling

Reference genome sequences of all species in this study were obtained from ancestral (i.e., unevolved) samples in the PacBio HiFi long-read sequencing laboratory (URL: https://www.pacb.com). Genomic sequences of each bacterial population sample were obtained using

short-read next–generation sequencing (NGS) equipment Illumina NovaSeq 6000. After assessing the read quality metrics of each prey population sample, genetic variants were obtained by following this read processing, variant calling and annotation workflow: Illumina DNA adapters trimmed with *cutadapt* (version v3.5; NBIS)[47], reads aligned to reference sequences with *BWA* (version 0.7.17; Li and Durbin)[48], read duplicates marked with *Picard* (version 2.27.5; Broad Institute, URL: https://broadinstitute.github.io/picard/), variants called in a multi-sample mode and subsequently filtered with *GATK Mutect2* (version 4.3.0.0; Broad Institute)[49], and finally annotated with *snpEff* (version 5.1d; Pablo Cingolani)[50]. Sequence data for populations coevolved with the predator extends beyond 800 days (up to approximately 1400 days) and all samples have been used to call the variants, however, because populations evolving alone were limited to a 800 day period, the variant data corresponding to samples taken later than 800 days into the experiment were not included in any downstream analyses of this study (see Supplementary Information for sample curation and tool usage details).

Clonal structure inference from mutation cohorts

Variant trajectories have been assigned to variant cohorts (clusters) using read counts of reference and alternate alleles with PyClone-VI[28]. Since this software is originally intended for human cancer genomics and the genomes under our study are haploid, we set the required parameters for major and minor copy numbers to one and zero respectively for every variant within the genomic time series data of any given population from our experiments. For each replicate separately, using 15 for an initial guess for the number of clusters with 100 restarts of the inference algorithm, we repeated the inference ten times, each with a different random seed (seed values 1, 2, ..., 10), and selected the model with the highest final evidence lower bound (ELBO) value, as described in the PyClone-VI manuscript[28]. The final number of inferred clusters varied from four to nine – however, we note that repeated inferences with different seed values may yield slightly different variant trajectory clusters and their numbers.

Given the computed variant cohorts of a prey population, we infer its clonal structure by following two clonality rules: 1) for two competing clones to co-exist, the sum of their frequencies cannot exceed 1 at any time point, and 2) the sum of the frequency of every sub-clone cannot exceed the frequency of the clone they belong to. The hierarchy of the clonal structure is represented by a tree, where each node represents a mutation cohort. As the cohorts "emerge" over time, i.e., their frequency exceeds a chosen threshold at a given time point (0.1 by default; some replicates needed higher thresholds, e.g., 0.15, to resolve obviously wrong hierarchical assignments due to multiple cohorts emerging at the same time with a similar frequency, close to the threshold), they are added as new nodes to the growing trees. After the first cohort has emerged and has been placed under the root node of the tree, every subsequent emerging cohort creates new trees, each representing one of the possible placements of the new node within each tree constructed before. Therefore, the total number of such possible trees that include every mutational cohort is the factorial of the number of inferred cohorts. The order of the cohort emergence is known, except in cases where two or more cohorts exceed the emergence threshold at the same time, doubling the number of possible trees with each additional emerging cohort at such a time point – because of that, we specify the order manually to prevent possibly degenerate cases due to incorrect clonality role assignment and to limit computational cost.

Next, having a set of complete clonal hierarchy trees for a given population, we use the mutational cohort frequencies Q to calculate the allele frequencies F that correspond to the inferred clonal structure at a given time t. Given a clone c and using $x \subset c$ to denote that clone x is a sub-clone of c, we define the allele fraction as:

$$F_c = Q_c - \sum_{x \subset c} Q_x$$

Since the originally computed mutation cohort frequencies are inherently noisy because of the variant trajectory variation within trajectory clusters, we use the newly calculated allele frequencies to marginalize the cohort frequencies, i.e., adjust their values per time point so that they fully agree with the two clonality rules. That is done by iteratively subtracting a small value (0.1% of the absolute F_c value) from the frequencies of alleles under consideration until their respective sum is either less than 1 or less than the frequency of the clone they are sub-clones of, respective to the placement of the considered allele in the previously inferred tree structure. This way, we evaluate the consistency of the inferred clonal hierarchy both quantitatively and visually.

Lastly, we select a tree with the lowest total marginalization cost to be the representative clonal structure of a given replicate. Since the marginalization of every possible tree is computationally time-prohibitive, we first select a subset of most likely tree configurations by implementing an arbitrary scoring function which corresponds to the two clonality rules; when considering a node that represents a cohort that emerged at time t, the score is the number of future time points $t_f > t$ where neither of the rules are violated, allowing up to 0.01 of absolute error to account for the potential noise in the frequency data. While the minimum score usually represents low-cost solutions, they may not be the lowest possible ones, and an objective solution other than the marginalization is not

known. To help mitigate this, we marginalize 50 best scoring clonal structures and choose the one with the minimal cost as our representative structure. Inferred clonal structures presented in this article are intended as examples of automated, yet realistic, solutions and not objectively best ones – it would likely be possible to find marginally lower-cost solutions.

We note that multiple rule-consistent and high scoring solutions are sometimes possible, especially when low-frequency cohorts emerge late in a population with multiple competing clones present at sufficiently high frequencies. Additionally, some cohorts never reach our emergence threshold, and we remove these cohorts altogether. Nevertheless, this method could be extended to potentially handle every possible configuration of the clonal structure tree in a more efficient way.

Phenotypic data

From each bacterial population in our study, in addition to four ancestral (day 0) isolates, ten isolates were taken at two time points in the experiment. The first sampling, early in the experiment, is done at days 135 (*S. marcescens*) and 64 (rest of the study species) from the evolved-alone populations; and at day 43 (all study species) from the coevolved populations. The second sampling, late in the experiment, is done at days 723 (*S. marcescens*) and 652 (rest of the study species) from the evolved-alone populations; and at day 631 (all study species) from the coevolved populations.

The isolates were then grown in three kinds of media: 5% King's Broth (KB) medium (without the presence of ciliate), predator medium (with ciliate) and salt stress medium (potassium chloride KCl; without ciliate). Growth curves were extracted by measuring the optical density of the grown bacterial colony every 20 minutes using the Agilent LogPhase 600 nm device. For each isolate, optical density measurements were replicated three times in total. Due to some inherent variability in these technical replicate measurements, we computed their median at each time point and used the resulting curve as our representative growth curve of that isolate.

Random forest classifier

To assess the classification accuracy in classifying phenotypic growth profiles (growth curves from all three growth media combined) according to species, evolutionary history, replicate and sampling time, we used scikit-learn[51] (version 1.4.2) Python (version 3.10.14) implementation of the random forest classifier to train 100 random forest models and averaged the resulting confusion matrices representing the test dataset over all model instances. Before learning each model, phenotypic growth profile data was randomly split in half for training and testing datasets in a stratified fashion, i.e., keeping the proportion of each unique label equal across both datasets. For result repeatability, each train-test split and random forest model instance were generated with a Numpy[52] (version 1.26.4) random seed $s = 0, \ldots, 99$ for each of the 100 runs.

Data availability

Raw sequence data (bam files) will be deposited in the EMBL ENA. Accession codes and links to be added upon acceptance of the manuscript.

Code availability

All code and pre-processed data needed to reproduce the downstream analyses and figures will be made available upon publication via GitHub: https://github.com/dovydask/clonal_structure.

Acknowledgments

We thank Liisa Ruusulehto, Paula Typpö, Roosa Jokela, Eeva Vakkari, Jutta Kasurinen, Iina Hepolehto, Tinja Hyvönen for their laboratory procedure work on the long-term experiment. We thank the WGGC Core Facility Cologne for the preparation and sequencing of our samples. We also thank Anthony Sun and Mikko Kivikoski for their help with downstream analyses and bioinformatics. This work was in part funded by the Research Council of Finland (Multidisciplinary Center of Excellence in Antimicrobial Resistance Research, grant no. 346126; grant nos. 330886 and 327741 to T.H; and grant no. 346128 to V.M.); Deutsche Forschungsgemeinschaft (BE 4135/12-1).

Author information

Authors and Affiliations

Department of Computer Science, Organismal and Evolutionary Biology Research Programme, University of Helsinki, Helsinki, Finland
Dovydas Kičiatovas, Johannes Cairns, Ville Mustonen

Department of Biology, University of Turku, Turku, Finland
Johannes Cairns, Teppo Hiltunen

Turku Collegium for Science, Medicine and Technology, Turku, Finland
Johannes Cairns

Department of Microbiology, University of Helsinki, Helsinki, Finland
Veera Partanen

Department of Biology, University of Konstanz, Konstanz, Germany
Julius Hoffmann, Lutz Becks

Contributions

J.C. and T.H. designed the long-term evolution experiment. J.C. and V.P. performed the long-term evolution experiment. J.H. performed phenotypic measurements. D.K. performed sequence data bioinformatics, downstream analyses and algorithm design. D.K., J.C., T.H., L.B. and V.M. were responsible for interpreting results. D.K. and J.C. wrote the original paper draft, with contributions from all authors. J.C. T.H., L.B. and V.M. supervised the work. All authors approved the final version of the paper.

Corresponding author

Correspondence to Lutz Becks, Teppo Hiltunen or Ville Mustonen.

Ethics declaration

Competing interests

The authors declare no competing interests.

References

1. Stern, D. L. *The genetic causes of convergent evolution* 2013.
2. Tenaillon, O. *et al.* The molecular diversity of adaptive convergence. *Science* **335.** ISSN: 10959203 (2012).
3. Wichman, H. A., Badgett, M. R., Scott, L. A., Boulianne, C. M. & Bull, J. J. Different trajectories of parallel evolution during viral adaptation. *Science* **285.** ISSN: 00368075 (1999).
4. Katz, S. *et al.* Dynamics of Adaptation during Three Years of Evolution under Long-Term Stationary Phase. *Molecular Biology and Evolution* **38.** ISSN: 15371719 (2021).
5. Papadopulos, A. S. *et al.* Rapid Parallel Adaptation to Anthropogenic Heavy Metal Pollution. *Molecular Biology and Evolution* **38.** ISSN: 15371719 (2021).
6. Saxer, G., Doebeli, M. & Travisano, M. The repeatability of adaptive radiation during long-term experimental evolution of Escherichia coli in a multiple nutrient environment. *PLoS ONE* **5.** ISSN: 19326203 (2010).
7. Bernardes, J. P. *et al.* The evolution of convex trade-offs enables the transition towards multicellularity. *Nature Communications* **12.** ISSN: 20411723 (2021).
8. Soudi, S. *et al.* Repeatability of adaptation in sunflowers reveals that genomic regions harbouring inversions also drive adaptation in species lacking an inversion. *eLife* **12.** ISSN: 2050084X (2023).
9. Maccagni, A. & Willi, Y. Trait divergence and trade-offs among Brassicaceae species differing in elevational distribution. *Evolution* **76.** ISSN: 15585646 (2022).
10. Lässig, M., Mustonen, V. & Walczak, A. M. Predicting evolution. *Nature Ecology and Evolution* **1.** ISSN: 2397334X (2017).
11. Frickel, J., Theodosiou, L. & Becks, L. Rapid evolution of hosts begets species diversity at the cost of intraspecific diversity. *Proceedings of the National Academy of Sciences of the United States of America* **114,** 11193–11198. ISSN: 10916490 (Oct. 2017).
12. Maharjan, R. P., Liu, B., Feng, L., Ferenci, T. & Wang, L. Simple phenotypic sweeps hide complex genetic changes in populations. *Genome Biology and Evolution* **7.** ISSN: 17596653 (2015).
13. Good, B. H., McDonald, M. J., Barrick, J. E., Lenski, R. E. & Desai, M. M. The dynamics of molecular evolution over 60,000 generations. *Nature* **551,** 45–50. ISSN: 0028-0836. https://www.nature.com/articles/nature24287 (Nov. 2017).
14. Lang, G. I. *et al.* Pervasive genetic hitchhiking and clonal interference in forty evolving yeast populations. *Nature* **500.** ISSN: 00280836 (2013).
15. Cairns, J., Jokela, R., Becks, L., Mustonen, V. & Hiltunen, T. Repeatable ecological dynamics govern the response of experimental communities to antibiotic pulse perturbation. *Nature Ecology and Evolution* **4.** ISSN: 2397334X (2020).
16. Corbett-Detig, R. B., Hartl, D. L. & Sackton, T. B. Natural Selection Constrains Neutral Diversity across A Wide Range of Species. *PLoS Biology* **13.** ISSN: 15457885 (2015).
17. Lion, S. Theoretical approaches in evolutionary ecology: Environmental feedback as a unifying perspective. *American Naturalist* **191.** ISSN: 15375323 (2018).
18. Retel, C. *et al.* The feedback between selection and demography shapes genomic diversity during coevolution. *Science Advances* **5.** ISSN: 23752548 (2019).
19. Harrison, E., Hall, J. P., Paterson, S., Spiers, A. J. & Brockhurst, M. A. Conflicting selection alters the trajectory of molecular evolution in a tripartite bacteria–plasmid–phage interaction. *Molecular Ecology* **26.** ISSN: 1365294X (2017).
20. Paterson, S. *et al.* Antagonistic coevolution accelerates molecular evolution. *Nature* **464.** ISSN: 00280836 (2010).
21. Lässig, M., Mustonen, V. & Nourmohammad, A. *Steering and controlling evolution — from bioengineering to fighting pathogens* 2023.
22. Johnson, M. S. *et al.* Phenotypic and molecular evolution across 10,000 generations in laboratory budding yeast populations. *eLife* **10.** ISSN: 2050084X (2021).
23. Cairns, J. *et al.* Evolution in interacting species alters predator life-history traits, behaviour and morphology in experimental microbial communities. *Proceedings of the Royal Society B: Biological Sciences* **287.** ISSN: 14712954 (June 2020).
24. Colclough, A., Corander, J., Sheppard, S. K., Bayliss, S. C. & Vos, M. Patterns of cross-resistance and collateral sensitivity between clinical antibiotics and natural antimicrobials. *Evolutionary Applications* **12.** ISSN: 17524571 (2019).
25. Scheuerl, T., Cairns, J., Becks, L. & Hiltunen, T. Predator coevolution and prey trait variability determine species coexistence. *Proceedings of the Royal Society B: Biological Sciences* **286.** ISSN: 14712954 (2019).
26. Lang, G. I. & Desai, M. M. The spectrum of adaptive mutations in experimental evolution. *Genomics* **104.** ISSN: 10898646 (2014).

27. Salcedo, A. *et al.* A community effort to create standards for evaluating tumor subclonal reconstruction. *Nature Biotechnology* **38.** ISSN: 15461696 (2020).

28. Gillis, S. & Roth, A. PyClone-VI: scalable inference of clonal population structures using whole genome data. *BMC Bioinformatics* **21,** 571. ISSN: 1471-2105. https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-020-03919-2 (Dec. 2020).

29. Dentro, S. C. *et al.* Characterizing genetic intratumor heterogeneity across 2,658 human cancer genomes. *Cell* **184.** ISSN: 10974172 (2021).

30. Frazão, N. *et al.* Two modes of evolution shape bacterial strain diversity in the mammalian gut for thousands of generations. *Nature Communications* **13.** ISSN: 20411723 (Dec. 2022).

31. Maddamsetti, R., Lenski, R. E. & Barrick, J. E. Adaptation, clonal interference, and frequency-dependent interactions in a long-term evolution experiment with escherichia coli. *Genetics* **200.** ISSN: 19432631 (2015).

32. Shao, K. T. & Sokal, R. R. Tree balance. *Systematic Zoology* **39.** ISSN: 1063-5157 (1990).

33. Rutherford, A. *Introducing Anova and Ancova: A GLM Approach* (2001).

34. Iwasaki, H. *et al.* A KaiC-interacting sensory histidine kinase, SasA, necessary to sustain robust circadian oscillation in cyanobacteria. *Cell* **101.** ISSN: 00928674 (2000).

35. Rao, N. N., Gómez-García, M. R. & Kornberg, A. *Inorganic polyphosphate: Essential for growth and survival* 2009.

36. Kimberly K Jefferson. What drives bacteria to produce a biofilm? *FEMS Microbiology Letters* **236,** 163–173. ISSN: 03781097. http://doi.wiley.com/10.1016/j.femsle.2004.06.005 (July 2004).

37. Ferrières, L. & Clarke, D. J. The RcsC sensor kinase is required for normal biofilm formation in Escherichia coli K-12 and controls the expression of a regulon in response to growth on a solid surface. *Molecular Microbiology* **50.** ISSN: 0950382X (2003).

38. Finn, R. D., Orlova, E. V., Gowen, B., Buck, M. & Van Heel, M. Escherichia coli RNA polymerase core and holoenzyme structures. *EMBO Journal* **19.** ISSN: 02614189 (2000).

39. Sule, P., Horne, S. M., Logue, C. M. & Prüb, B. M. Regulation of cell division, biofilm formation, and virulence by FlhC in Escherichia coli O157:H7 grown on meat. *Applied and Environmental Microbiology* **77.** ISSN: 00992240 (2011).

40. Mollet, C., Drancourt, M. & Raoult, D. rpoB sequence analysis as a novel basis for bacterial identification. *Molecular Microbiology* **26.** ISSN: 0950382X (1997).

41. Hiltunen, T. *et al.* Dual-stressor selection alters eco-evolutionary dynamics in experimental communities. *Nature Ecology and Evolution* **2,** 1974–1981. ISSN: 2397334X (Dec. 2018).

42. Meroz, N. *et al.* Evolution in microbial microcosms is highly parallel, regardless of the presence of interacting species. *Cell Systems* **15,** 930–940. ISSN: 24054712 (Oct. 2024).

43. Strelkowa, N. & Lässig, M. Clonal interference in the evolution of influenza. *Genetics* **192.** ISSN: 00166731 (2012).

44. Łuksza, M. & Lässig, M. A predictive fitness model for influenza. *Nature* **507.** ISSN: 14764687 (2014).

45. Becks, L., Hilker, F. M., Malchow, H., Jürgens, K. & Arndt, H. Experimental demonstration of chaos in a microbial food web. *Nature* **435,** 1226–1229. ISSN: 00280836 (June 2005).

46. Matz, C. & Kjelleberg, S. Off the hook - How bacteria survive protozoan grazing. *Trends in Microbiology* **13.** ISSN: 0966842X (2005).

47. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17** (2011).

48. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25.** ISSN: 13674803 (2009).

49. Benjamin, D. *et al.* Calling Somatic SNVs and Indels with Mutect2. *bioRxiv* (2019).

50. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly* **6.** ISSN: 1933-6934 (2012).

51. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **12.** ISSN: 15324435 (2011).

52. Harris, C. R. *et al.* *Array programming with NumPy* 2020.