Cryptic genetic variation promotes rapid evolutionary adaptation in an RNA enzyme

Hayden, E J ; Ferrada, E ; Wagner, A

Abstract: Cryptic variation is caused by the robustness of phenotypes to mutations. Cryptic variation has no effect on phenotypes in a given genetic or environmental background, but it can have effects after mutations or environmental change. Because evolutionary adaptation by natural selection requires phenotypic variation, phenotypically revealed cryptic genetic variation may facilitate evolutionary adaptation. This is possible if the cryptic variation happens to be pre-adapted, or "exapted", to a new environment, and is thus advantageous once revealed. However, this facilitating role for cryptic variation has not been proven, partly because most pertinent work focuses on complex phenotypes of whole organisms whose genetic basis is incompletely understood. Here we show that populations of RNA enzymes with accumulated cryptic variation adapt more rapidly to a new substrate than a population without cryptic variation. A detailed analysis of our evolving RNA populations in genotype space shows that cryptic variation allows a population to explore new genotypes that become adaptive only in a new environment. Our observations show that cryptic variation contains new genotypes pre-adapted to a changed environment. Our results highlight the positive role that robustness and epistasis can have in adaptive evolution.

Cryptic genetic variation promotes rapid evolutionary adaptation in an RNA enzyme.

Eric J. Hayden, Evandro Ferrada, Andreas Wagner

Institute of Biochemistry, University of Zurich, Zurich, Switzerland.

Cryptic genetic variation has no effects on phenotypes; such effects become manifest only after mutations or environmental change. Caused by the robustness of phenotypes to mutations, cryptic genetic variation has important implications for the study of disease, for animal and plant breeding, and for the evolution of novel traits in natural populations. Whether it facilitates evolutionary adaptations (and why) has remained elusive, partly because most pertinent work focuses on complex phenotypes of whole organisms whose genetic basis is incompletely understood. We here study cryptic variation in a simpler phenotype, the catalytic activity of an RNA enzyme. We find that populations of RNA enzymes with accumulated cryptic variation adapt more rapidly to a new chemical environment than a population without cryptic mutations. A detailed analysis of our evolving RNA populations in genotype space shows that cryptic variation allows a population to explore new genotypes that become adaptive only in a new environment. Our observations show that cryptic variation contains new genotypes pre-adapted to a changed environment. It highlights the positive role that robustness and epistasis can play in adaptive evolution.

The existence of cryptic genetic variation has been demonstrated in several organisms by the broad range of abnormal phenotypes that result from environmental or mutational perturbations[1]. Cryptic variation is only conditionally neutral to selection. It is therefore a distinct class of neutral mutations, where neutrality is dependent upon environmental and genetic context. Classic examples come form the production of phenocopies in *Drosophila*. Phenocopies are abnormal phenotypes that resemble the effect of major mutations in specific genes. However, phenocopies occur in stressful environments in individuals that are wild-type at these genes. This is possible because wild populations of *Drosophila* that are highly uniform in morphology are still genetically diverse. An average of six nucleotide differences per thousand base pairs (in *D. melanogaster*) means that each fly in an outbred population is genetically unique[2]. Phenocopies are caused by the exposure of this cryptic variation under "stressful" conditions. The genetic basis of phenocopies was confirmed through artificial selection[3]. This example and many others demonstrate that genetic variation is common, but its expression as phenotypic variation is highly buffered, and dependent upon environmental and genetic context[4].

The hiding of genetic variation requires a phenotype's robustness against the effect of mutations, a phenomenon which is often called mutational robustness[5]. Mutational robustness does not require a complex phenotype, and is biologically important even at the level of individual protein and RNA molecules. A given RNA or protein molecule can accept many mutations without changing structure or losing function[6,7]. As a consequence of this robustness, many different sequences in a genotype space have the same phenotype. They form the same tertiary structure and have the same biochemical activity. These sequences form vast mutational networks whose members

can be connected through series of mutations in their nucleotides. They are thus accessible from each other through small evolutionary changes. These mutational networks are also sometimes called "genotype networks" or "neutral networks" because they consist of different genetic sequences that have similar phenotypes[8,9]. RNA has been used as a model system for the study of genotype networks both computationally and experimentally[10-14], because it provides accessibility of both genotypic (sequence) and phenotypic (structure/activity) information. These studies confirm the existence of vast mutational networks in RNA structures that indicate a fundamental mutational robustness of RNA phenotypes[15], which results in the potential for cryptic variation at the molecular level.

Evolutionary adaptation by natural selection requires phenotypic variation. Phenotypically revealed cryptic genetic variation may thus facilitate evolutionary adaptation[16-18]. This is possible if the cryptic variation happens to be pre-adapted[19] or "exapted"[20] to the new environment, and is thus advantageous once revealed. However, this facilitating role has not been proven. Part of the reason is that cryptic variation has traditionally been studied for complex macroscopic traits. The link between genotype and phenotype is poorly understood for such traits, and it is difficult to study cryptic variation systematically for them. We decided to undertake the study of cryptic genetic variation in a simpler system, a catalytic RNA molecule, where we can access both phenotypic and genotypic information. Further, our experimental system allows us to quantify and visualize genetic variation during evolutionary adaptation using next generation sequencing technology.

For our evolution experiments, we chose the well-characterized group I RNA enzyme (ribozyme) derived from the *Azoarcus* pre-tRNA[ile]. We chose this ribozyme primarily for its robust phenotype. The *Azoarcus* ribozyme folds consistently and rapidly (< 100 ms) to the active state without getting trapped in misfolded intermediates, which is the fate of many large RNA molecules, including other group I ribozymes[21,22]. Once reached, the active state is highly stable, and retains activity at unusually high temperatures (80 °C), or in the presence of high concentrations of denaturants (7.5 M urea)[23]. We thus expected the robust phenotype of the *Azoarcus* ribozyme to tolerate mutations without losing function, making it an ideal candidate for the experimental study of cryptic variation. In addition, the *Azoarcus* ribozyme has been the subject of several studies focusing on ribozyme folding and kinetics[22,24], and a high resolution crystal structure of the ribozyme in an active conformation has been solved[25]. These detailed functional and structural data provide guidance for designing and interpreting *in vitro* evolution experiments.

We utilized a modified version of a previously reported procedure for the *in vitro* evolution of group I introns[26,27] (Fig. 1a). The procedure begins with a population of $\sim 10^{13}$ variant ribozymes generated by mutagenesis of the wild-type sequence (Methods). This population is challenged to catalyze the sequence specific cleavage of an exogenous nucleic acid substrate. As a by-product of this cleavage reaction, a portion of the substrate becomes ligated to the 3'-end of the ribozyme, which generates a primer binding site for specific amplification of molecules only if they have successfully performed a catalytic event. The random mutations that are introduced during the amplification process, along with selection for catalytic activity, allow the system to evolve based on Darwinian principles. At each "generation" (one cycle of mutagenesis and selective-amplification) the activity of the population can be

determined at the ribozyme reaction step, and a cDNA library is produced which can be characterized to monitor genetic changes.

To produce populations containing cryptic variation, we exposed populations of RNA enzymes to a period of mutagenesis with purifying selection, by selecting for molecules with the native RNA cleavage activity (Fig. 1b). Thus, by continuously introducing new variants through random mutation, while also selecting for activity on a "native" RNA substrate, we expected to accumulate mutations in the evolving populations while maintaining the native ribozyme function. We carried out two independent evolution experiments, and called them "line A" and "line B". They were identical except for the addition of 5 M formamide to line B. Formamide is a denaturant that lowers the melting temperature of base paired regions by approximately 2.7 ^{o}C/M[28]. Thus, the addition of formamide in line B introduces a more stringent selection pressure, which in terms of structural stability, is analogous to increasing the selection temperature by 13.5 ^{o}C, but without the numerous other consequences (i.e. on pH and k_{cat}) of a temperature change. Pilot experiments showed that the addition of formamide lowered the amount of ribozyme reacted under the selection conditions by about 30% (Supplementary Fig. 1). We were concerned by the possibility that the weaker selection conditions of line A could allow the accumulation of moderately deleterious mutations, causing a continuous decline in the population's activity. Also, ribozymes selected in the laboratory may have a lower mutational robustness than natural ribozymes[29]. The inclusion of formamide in line B was intended to counteract these possibilities. However, as we shall see, neither line experienced a decrease in the native activity while accumulating genetic variation.

We used the same starting population for each line. We constructed this population by mutagenesis of 159 nucleotides of the wild-type *Azoarcus* ribozyme sequence through a mutagenic PCR[30]. The diversity of this starting population can be calculated by determining the probability of finding sequences with a given numbers of mutations from the "wild-type" sequence using binomial statistics (Methods). Based on an estimated mutation rate of 0.0066 per nucleotide per PCR, individual sequences in the population contained on average 1 mutation relative to the wild-type. Our large population (~10^{12} individuals; 160 ng PCR produced dsDNA) also contained all possible sequences with four or fewer nucleotide changes, and random samples of sequences with five or more changes relative to the wild-type sequence.

We monitored the native activity of the two lines over 10 generations of purifying selection while introducing mutations at every generation (Fig. 2a). No significant difference in activity (two-tailed t-test, 95% C.I.) exists between the initial and final round of selection for both lines. This demonstrates that the procedure was in fact maintaining the native activity of the population despite the high mutation rate, and provides evidence supporting the cryptic nature of any accumulated mutations. Although the activity appears to decrease to a minimum at generation four, followed by a recovery, no significant difference in activity exists between any two generations within either lineage. Line B, with higher selection stringency, preserved the same level of activity as line A.

Although both lines maintained the initial activity, sequencing analysis revealed that mutations had in fact accumulated. We determined the DNA sequences of 2748 ± 770 molecules sampled from each generation. We first measured the composite genetic

change of the population by determining the mutational distance between individuals in the populations and the original wild-type sequence (Fig. 2b and c). That mutations accumulated in both lines was indicated by the mean increase in population distance from the wild-type sequence over time (Fig. 2b and Supplementary Fig. 2). At every generation, line B showed a slightly higher mean distance to the wild-type than line A, although divergence in the last generations was similar (Fig. 2b). While both populations maintain some copies of the wild-type sequence, by generation 10 fewer than 1% of the sampled individuals had no mutations (Supplementary Fig. 2). To identify the positions where the mutations accumulated, we classified positions as either "mutable" or "non-mutable" (Fig. 2c and d). Mutable positions showed a rate of mutational change significantly greater than zero, based on linear regression (Methods). We found that 35 positions (22%) in line A, and 19 positions (12%) in line B were mutable, with 15 of these positions common to both lineages.

Although the accumulated mutations did not affect the phenotype of the population under "native" conditions, we hypothesized that they could facilitate evolutionary adaptation of the populations to a new chemical environment. To test this hypothesis, we challenged the resulting populations to adapt to a non-native function by changing the substrate in the selection procedure. For this, we chose an RNA substrate with identical sequence, but with a phosphorothioate (PT) replacing the scissile phosphate (Fig. 1b). This chemical change represents a "promiscuous activity"[31] of the *Azoarcus* ribozyme with a 200% decreased catalytic efficiency (k_{cat}/Km), by mostly affecting k_{cat}[24]. We started new evolution lines from ~10^{13} (20 pmol) RNA molecules taken from the last generation of line A and line B. We called these lines *New*-A and *New*-B, respectively. In addition, we started another new line from a sample of ~10^{13} RNA

molecules taken from the original initial population. We called this line *New*-wt. In this phase of the experiment we wanted to analyze the effect of previously accumulated mutations on evolutionary adaptation to the new substrate. Thus, we used the same reaction conditions for all three lines, and reduced the mutation rate by several-fold to approximately 0.001 per nucleotide per generation by replacing the mutagenic PCR step of our selection procedure with a standard PCR (Methods). This lowered mutation rate favors error free replication of the reacted ribozymes, i.e. 98% of sequences will be replicated without mutation. However, at every generation, statistics predicts that the population will still experience every combination of newly introduced mutations involving one or two mutated positions (Methods).

We selected for activity with the new substrate during eight generations, and again measured the activity of each population at each generation as the amount of ribozyme reacted (Fig. 3a). In each line, the activity increased significantly between the first and last generations. However, lines *New*-A and *New*-B showed a much faster rate of adaptation than line *New*-wt. To facilitate comparison, we calculated the rate of adaptation by dividing the percent increase in activity (fraction ribozyme reacted) by time (generations). The greatest difference in rate is found at generation 5, where the rates of adaptation for lines *New*-A, *New*-B and *New*-wt were 19.5, 15.5, and 2.5, respectively. This corresponds to an approximately 8-fold faster rate of evolutionary adaptation for line *New*-A relative to line *New*-wt, and an approximately 6-fold faster rate for line *New*-B.

We analyzed the genotypic changes in the new evolving lines at varying levels of detail. We first measured the composite genetic change by measuring the distance

between individuals in the populations and the wild-type sequence (Fig. 3b). Surprisingly, all three lineages showed different diversity, even though they had evolved under the same conditions. Line *New*-A had a very high mean distance from wild-type. This mean distance further increased over time from 10.6 to 13.7 nucleotide differences. Line *New*-B showed an intermediate mean distance that increases slightly from 7.7 to 9.1. Line *New*-wt showed the lowest distance values, but it also increased from 2.8 to 4.4 nucleotide differences.

We next identified genotypes that were potentially contributing to the increasing activities of the evolving populations. To this end we defined genotypes by their unique combination of mutations relative to the wild-type. We determined the population frequency of different combinations of mutations, and identified genotypes that increased most rapidly in their frequency with respect to generation time. We limited our search to combinations of individual mutations that showed the most significant increase in frequency with time (Fig. 3c). Significance was determined by linear regression with correction for multiple testing (Supplementary Fig. 3 and Methods). It should be pointed out that the increase in the frequency of genotypes that are under selection is expected to be exponential, not linear, under standard population genetic models[32]. Linear regression penalizes non-linearity, and thus our determination of significance of individual mutations is a conservative approach.

Two important genotypes stand out (Fig. 3d). In line *New*-A, the most rapidly increasing genotype, which we call *AzoΔ*, includes deletions at positions 47-53 combined with seven point mutations (G31U, G35U, G70U, G121A, C141U, A144G and G183C). By generation 8, this genotype represented 31% of the population, and

subsets of its mutations accounted for 56% of the population. In line *New*-B, the most rapidly increasing genotype, which we call *Azo**, is composed of four point mutations (G32U, G53A, C89U and G179C). By generation 8, sequences containing all four mutations accounted for 23% of the population, and various subsets of these four mutations accounted for 78% of the population. To corroborate the significance of association between the individual mutations of these genotypes, we also calculated correlation coefficients[32] for all pairs of mutations for either *AzoΔ* or *Azo** (table S1). All pairs showed significant correlation (P < 0.05, chi-squared), supporting the conclusion that these groups of mutations represent individual genotypes.

One can estimate the fitness *w* of a genotype relative to the rest of a population from its increase in frequency over time according to the expression $\ln(p_t/q_t) = t\ln(w) + \ln(p_0/q_0)$, where p_t and q_t are the frequencies of this genotype and of all other genotypes, respectively, at time t[32]. This calculation yields a relative fitness for *AzoΔ* of $w = 1.46$, or a 46% advantage over the remainder of the population. The genotype *Azo** has a fitness of $w = 1.82$, or an 82% advantage.

The fitness advantage for a sequence could potentially arise at any step in our selection procedure, such as during reverse transcription or PCR. To determine if the high fitness genotypes were in fact important contributors to the increased activity observed in Fig. 3a, we tested whether they had an increase in activity with a phosphorothioate substrate. To this end, we synthesized clonal transcripts of the *AzoΔ*, *Azo**, and *wild-type* ribozymes for kinetic analysis. We performed the kinetic experiments under the same conditions as used in the *New* selection lines, and

monitored the production of the selected species, i.e. the 3'-modified ribozyme
(Methods).

Surprisingly, the clonal preparation of the *AzoΔ* ribozyme showed no activity toward
the selected PT substrate. We further tested two other ribozyme variants, one with
only the deletion mutations, and another with only the seven point mutations. These
variants also showed no activity with the PT substrate, demonstrating that either group
of mutations is enough to obliterate activity. Thus, not only did this genotype survive
our selection procedure without any individual activity, it actually showed a dramatic
increase in frequency, suggesting a selective advantage within the population.
Because our selection procedure specifically requires the production of a 3'-modified
ribozyme, we needed to account for the selection of a molecule with no apparent
activity. We hypothesized that this sequence lacked the ability to fold into the native
state *individually*, but could form an active complex in conjunction with other active
ribozymes. Such an intermolecular partnership was previously observed in several
other ribozyme experients[33-35]. To test this hypothesis, we assayed the *AzoΔ* ribozyme
for activity with a PT substrate alone, or in the presence of either the *wild-type*
Azoarcus or the *Azo** ribozymes (Fig. 3e). In these experiments, only the *AzoΔ*
ribozyme was 5'-radiolabled with ^{32}P so that only the activity of this ribozyme was
observable on a denaturing polyacrylamide gel. The results confirm that while the
AzoΔ ribozyme is inactive individually, it regains activity upon addition of either
active variant.

For the *Azo** ribozyme, we found an increased activity with a phosphorothioate
substrate as compared to the *wild-type Azoarcus* ribozyme (Fig. 3f). The four

mutations of *Azo** increase the observed rate constant (k_{obs}) by 131%, and also

increase the extent of ribozyme reacted by 76%. Thus, the presence of this genotype

in the population accounts for much of the increasing activity of line *New*-B.

Because the increased activity of the *Azo** variant correlates with increased activity of

line *New*-B, we also looked for the presence of the *Azo** genotype in lines *New*-A and

New-wt. In line *New*-A, *Azo** is present and increases in frequency from 1.1% to 8.0%

over eight generations, which corresponds to a fitness of $w = 1.35$. Thus, this variant

also appeared in this line, but the fitness of this genotype is lower here than in line

New-B. The lowered fitness is presumably a consequence of the presence of

individually inactive, yet highly fit *AzoΔ* variants. Because the increase in the

frequency of *Azo** is modest in this line, linear regression had not identified the four

mutations as individually significant, which demonstrates a limitation of this

approach.

In line *New*-wt, the *Azo** genotype did not appear in the first three generations, and

only three individuals contained all four mutations in generation 8 (0.2% of sample).

The most frequent combination of the individual *Azo** mutations was G53A and

G179C, which accounted for 9.6% of the population by generation 8 (Fig. 3d). We

confirmed the absence of significant association of these mutations by calculating

correlation coefficients for each possible pair of mutations (table S1). For example,

while the individual mutations G53A and G179C were quite frequent by generation 8

(each >30% of the sampled individuals), they co-occurred at only 5% of the

theoretical maximum (Methods). This suggests that the significant individual

mutations from line *New*-wt are acting separately, not in concert. Further, the other

individual mutations of *Azo** (C32U and C89U) did not show a significant increase in frequency in this line, as determined by linear regression, suggesting that they do note display a strong selective advantage individually. In sum, although the *Azo** genotype showed increased activity with the phosphorothioate substrate, line *New*-wt, which had not yet acquired cryptic variation, did not discover this genotype.

While the *Azo** genotype has an increased activity in the new environment, our data suggest that the mutations that comprise this genotype had no advantage in the native environment. First, the composite activities of the populations of lines A and B did not increase during selection for the native activity, suggesting that these lines had not yet discovered higher fitness genotypes. Also, the individual mutations of the *Azo** genotype do not increase in frequency during selection for the native activity (Supplementary Fig. 4), and the *Azo** genotype was not detected in line A or B.

To confirm the cryptic nature of the *Azo** mutations, we engineered these mutations into the *Azoarcus* ribozyme. We then determined the activity of these variants under the conditions used during selection for the native activity in line A (no formamide) and compared them to the wild-type activity (Fig. 4a). Of the individual mutations of the *Azo** genotype, only G179C showed an increase in mean activity (14%), but which was not significantly different from the wild-type ($p = 0.11$, two-tailed t-test). The mutation C32U showed a significant decrease in activity (-33%, $p = 0.03$). The mutations G53A and C89U both showed decreased in activities (-28% and -17% respectively) that were not significantly different than the wild-type ($p = 0.07$ and $p = 0.10$, respectively). Thus, the individual mutations of the Azo* genotype showed no fitness benefit during selection for the native activity. However, because three of these

individual mutations, and several combinations (Supplementary Fig. 5), showed no significant difference from the wild-type, they can be discovered and remain in the population despite purifying selection for the native activity. This is consistent with the observations that approximately 10% of sampled individuals in generation 10 of lines A and B had at least one of the *Azo** mutations, but none of the mutations showed a dramatic increase in frequency.

Further, due to the diversity of our populations, the *Azo** mutations rarely occurred in an entirely wild-type background. For example, after the first round of selection for the native activity, only ~4% of line A individuals and ~6% of line B individuals are wild-type, and there are almost no wild-type individuals left in either line by generation 10 (Supplementary Fig. 2). Also, most selected individuals will acquire a random mutation during amplification. Thus, we wanted to test the average effect of mutations on the *Azo** genotype. To this end, we introduced random mutations into the *Azo** genotype through mutagenic PCR (Methods), and tested the activity of this population of ribozymes on either the native or new substrate (Fig. 4b). The results showed that in the native environment, the effect of random mutations decrease the average activity of *Azo** individuals to approximately the same level as the wild-type activity ($p = 0.91$, two-tailed t-test). However, in the new environment, the activity of the population of *Azo** mutants remained significantly higher than the wild-type activity despite random mutations ($p = 0.007$).

We conclude that the mutations of the *Azo** genotype represent cryptic variation that is advantageous in the new environment, but had no distinguishable selective

advantage in the native environment. The lack of selective advantage is compounded by the average deleterious effect of random mutations.

We next turn to a more detailed visual analysis of sequence space to help us understand why cryptic variation allowed faster adaptation. This space is very high-dimensional and cannot be visualized directly. However, we can study lower-dimensional projections of this space, for example using principal component analysis (PCA) of aligned sequence data sampled from evolving populations. Fig. 5 shows such an analysis based on sequences isolated from three generations of the *New* lines. It shows that, first, lines *New-A* and *New-B* are more diverse during all generations, compared to line *New*-wt. Second, it also confirms the existence of two subpopulations of line *New-A,* where two clearly discernible clouds of sequences are visible at all times; one contains the *Azo** genotype, and the other contains *Azo⊿*. Third, it illustrates the high fitness of *Azo⊿* and *Azo** in that the number of sequences belonging to these genotypes increases over generational time. Importantly, it shows that many of the sequences in generation 1 of lines *New*-A and *New*-B are close in genotype space to *Azo**. Over time, the genotypes become more concentrated around the *Azo** genotype. In contrast, in generation 1 of line *New*-wt, sequences are tightly clustered and distant from *Azo**. Over time, this population becomes more diverse, and moves towards the region of space occupied by the *Azo** individuals.

A candidate explanation for the advantage of cryptic variation that emerges from the previous analyses is that lines A and B had the opportunity to expand in sequence space, such that their sequences came close to regions where advantageous mutations could occur in line *New*-A and *New*-B. Line *New*-wt did not have this opportunity,

and thus adapts more slowly. Thus, while the genetic variation acquired during purifying selection did not affect the activity on the native substrate, it allowed the rapid adaptation after the environmental perturbation. This rapid adaptation coincides with the discovery of *Azo**, a variant with increased activity. To further validate this explanation, we constructed a network of sequences combined from the last generation of line B (B10) and the first generations of lines *New*-B (*New*-B1) and *New*-wt (*New*-wt1) (Fig. 6). In Fig. 6a, this network is represented as a graph, whose nodes correspond to clusters of sequences (96% identity) and edges connect related clusters (distance ≤10 nucleotide changes). The graph shows that many of the individuals in B10 (7 clusters) are already in close proximity (connected by a yellow edge) to the high fitness *Azo** genotype (yellow nodes with black borders). The first round of selection in line *New*-B (*New*-B1) acts on this standing diversity and increases the number of sequences (33 clusters) around *Azo**. In obvious contrast, not a single cluster from *New*-wt1 is in close proximity to an *Azo** cluster. The proximity of B10 individuals to *Azo**, relative to *New*-wt1 individuals, is further supported by analyzing the positions where the *Azo** mutations occur: 32, 53, 89, and 179 (Fig. 6 b-d). The results show that many sequences in B10 already posses two or three of the four *Azo** mutations. No individuals in *New*-wt1 posses three *Azo** mutations, and only a fraction of a percent posses two. Thus, the cryptic diversity acquired during purifying selection for the native activity moved some of the population to regions of genotype space that happen to be proximal to the high fitness genotype.

To visualize *all* the sequence data from these three generations requires clustering which compromises resolution. We thus constructed an unrooted phylogenetic tree to understand the relatedness between *Azo** sequences and closely related sequences in

B10 at high resolution (Fig. 6e). To this end we used only the sequences that share more than 96% identity with *Azo**, which are those found in the clusters represented as yellow nodes in the network graph of Fig. 6a (Methods). For comparison, we also included the representative sequences form *New*-wt1, although they do not share this level of sequence identity with *Azo**. The tree shows, first, that at this early stage of evolution toward the *New* activity, the sequences that contain the *Azo** mutations are not spread throughout the tree, but are related, and form a relatively small clade. Second, it shows that the sequences of B10 are approximately equally distant from the different clades in *New*-B1, including the "*Azo** clade", which arise from the first round of selection for the new activity. Third, sequences from *New*-wt1 are much more distant to these *Azo** sequences. The tree supports the notion that the sequences with cryptic mutations in B10 are not specifically adapted for the new substrate activity because they are equally close to clades that do not contain *Azo**. Nevertheless, these sequences are fortuitously similar enough to the *Azo** sequences to allow the immediate emergence of this high fitness genotype upon environmental change (new substrate). This was not the case in line *New*-wt, which lacked this cryptic variation.

In summary, our results show that populations which had acquired cryptic variation adapted rapidly to a chemically perturbed substrate. Their rapid adaptation results from the presence of a high fitness genotype that requires four mutations relative to the wild-type sequence. The required mutations are cryptic when selecting for the native activity; they do not display a selective advantage. However, upon environmental perturbation (new substrate), the high fitness phenotype of these genotypes is revealed, and selected, causing rapid adaptation. We also observed the

origin of a simple molecular "ecology" in the interactions between the *AzoΔ* variant and active enzymes, showing how unforeseen complexity can arise in seemingly simple systems.

Our observations demonstrate that cryptic variation can facilitate adaptation, and why it does so. Populations under purifying selection for a trait can still evolve genotypic diversity, because there are many different genotypes with the same or similarly well-adapted phenotype. Some of this diversity is fortuitously pre-adapted to a new environment, which aids the population's evolutionary adaptation to this new environment. We note that this genotypic diversity is a signature of extensive epistasis. Indeed, such epistasis has recently been demonstrated in protein and RNA phenotypes[36-38]. Our results suggest that epistasis is important in our system, because several individual mutations do not provide a large fitness advantage alone, but do so in combination. The ability to explore such combination of mutations cryptically is especially important in cases where high-fitness genotypes require several interacting mutations. Our observations also support theoretical work which demonstrates that the release of hidden variation after perturbation is a general property of genetic systems near mutation-selection balance, and with epistatic or gene-environment interactions[39].

Our results also demonstrate the importance of robustness in facilitating evolutionary adaptation[40,41]. Cryptic variation becomes possible only because the enzymes we study are to some extent robust to genetic change, and can explore a large genotype space even when subject to purifying selection. Our more rapidly adapting populations are also more diverse, providing experimental evidence of how

mutational robustness can lead to the fulfillment of Fischer's fundamental theorem[42]. Strikingly, the robustness to mutations of the *Azo** genotype in the native and new environments provides a selective advantage for these individuals only after environmental change (new substrate). We note that this different robustness could be the result of either a buffering of deleterious mutations or from a high density of more active sequences in the local genotype space[38].

The phenotype of our study system is much simpler than complex traits of higher organisms, which is its primary limitation. However, this system allows the monitoring of population wide genotypic change over multiple generations through ultra-high throughput sequencing. This ability allowed us to analyze the relationship between standing variation and high fitness genotypes over time, and to quantify the *rate* at which combinations of mutations rise to high frequency. Our results suggest that we may understand the role of cryptic variation in complex traits to the extent that we can analyze their evolution in an underlying genotype space.

Our observations suggest that the adaptive advantage of cryptic variation will often be transient. Unless the individual mutations that lead to a high fitness genotype are strongly deleterious, even a population without cryptic variation will eventually discover them, and climb the fitness landscape to this genotype. This transient effect is evident in Fig. 2b, where the three *New* lines are approaching a common activity. We note that our system involved only a single environmental change, whereas in the wild, environments change incessantly. Even if a robust system with cryptic variation only has a transient advantage after environmental change, this advantage may become permanent if environments change frequently. Consistent with this notion is

that protein structures with greater robustness to mutations and thus greater potential for cryptic variation have evolved greater functional diversity in their evolutionary history[43].

Our results are similar to those found in a protein system, where mutations that did not affect the native activity of the protein changed the activity of non-native or "promiscuous" functions, and in some cases were several mutations closer to new phenotypes[44]. Thus, our results support a general role for cryptic variation in accelerating adaptive evolution. Further, if cryptic variation is important in simple, single enzyme systems, it is likely to be pervasive in complex phenotypic traits. The ability to sequence entire genomes and evolving populations through deep sequencing is leading to a surge in genome based research. These efforts are continually uncovering single-nucleotide polymorphisms that have yet unknown effects, and might often be hidden under normal conditions. Our observations suggest that these mutations should not be ignored, and may have important consequences in populations that face perturbations, including new mutations, infectious disease agents, and environmental changes.

METHODS SUMMARY

The dsDNA template for the *Azoarcus* ribozyme was produced from a two step PCR based assembly of synthetic oligonucleotides[45]. Ribozyme populations were prepared from *in vitro* transcritption (T7 RNA polymerase) and purified for length homogeneity by denatureing PAGE (6% polyacrylamide/8M urea). Mutagenesis was achieved by a mutagenic PCR procedure[24], and to a lesser extent by the inherent mutation rates of the polymerase enzymes of the selection procedure. Substrate

oligonucleotides were produced by solid phase synthesis and purified by PAGE

(Microsynth). Selection was based on a reverse splicing reaction containing 20 pmol

ribozymes, and either 100 pmol RNA oligonucleotide substrate, or 200 pmol

phosphorothioate containing substrate (equal mixture Rp/Sp). Negative controls for

the selection protocol were carried out for every generation by skipping the reverse

transcription step, but keeping the remainder of the protocol identical, and were

monitored at both PCR steps by agarose gel electrophoresis. No band was ever

observed in a negative control. Kinetic parameters were determined by non-linear

curve fitting of time course data (Methods). cDNA samples from each generation

were appended with a primer sequence unique to that generation via a PCR reaction.

Samples from all generations were combined, and sequenced on a single picotiter

plate using Roche 454 technology by the Functional Genomics Facility Zurich. P-

values from linear regression were adjusted for multiple testing using the Benjamini

Hochberg procedure[46]. Principle component analysis was performed using the

princomp function in Matlab on multiple sequence alignments constructed from

combined generations. Network graphs were constructed using Cytoscape from all-

against-all distance data extracted from the multiple sequence alignments.

1. Gibson, G. & Dworkin, I. Uncovering cryptic genetic variation. *Nat Rev Genet* **5**, 681-690 (2004).
2. Powell, J.R. *Progress and prospects in evolutionary biology: the Drosophila model*. (Oxford University Press: 1997).
3. Waddington, C.H. Genetic Assimilation of an Acquired Character. *Evolution* **7**, 118-126 (1953).
4. Rutherford, S.L. From genotype to phenotype: buffering mechanisms and the storage of genetic information. *Bioessays* **22**, 1095-1105 (2000).
5. Wagner, A. *Robustness and Evolvability in Living Systems*. (Princton University Press: 2005).
6. Rennell, D., Bouvier, S.E., Hardy, L.W. & Poteete, A.R. Systematic mutation of bacteriophage T4 lysozyme. *Journal of Molecular Biology* **222**, 67-86 (1991).
7. Ancel, L.W. & Fontana, W. Plasticity, evolvability, and modularity in RNA. *J. Exp. Zool.* **288**, 242-283 (2000).

8. Schuster, P., Fontana, W., Stadler, P.F. & Hofacker, I.L. From Sequences to Shapes and Back: A Case Study in RNA Secondary Structures. *Proceedings: Biological Sciences* **255**, 279-284 (1994).

9. Wilkins, A.S. Colloquium Papers: Between "design" and "bricolage": Genetic networks, levels of selection, and adaptive evolution. *Proceedings of the National Academy of Sciences* **104**, 8590-8596 (2007).

10. Carothers, J.M., Oestreich, S.C., Davis, J.H. & Szostak, J.W. Informational Complexity and Functional Activity of RNA Structures. *Journal of the American Chemical Society* **126**, 5130-5137 (2004).

11. Lehman, N., Donne, M.D., West, M. & Dewey, T.G. The Genotypic Landscape During In Vitro Evolution of a Catalytic RNA: Implications for Phenotypic Buffering. *Journal of Molecular Evolution* **50**, 481-490 (2000).

12. Held, D.M., Travis Greathouse, S., Agrawal, A. & Burke, D.H. Evolutionary Landscapes for the Acquisition of New Ligand Recognition by RNA Aptamers. *Journal of Molecular Evolution* **57**, 299-308 (2003).

13. Lincoln, T.A. & Joyce, G.F. Self-Sustained Replication of an RNA Enzyme. *Science* **323**, 1229-1232 (2009).

14. Schultes, E.A. & Bartel, D.P. One Sequence, Two Ribozymes: Implications for the Emergence of New Ribozyme Folds. *Science* **289**, 448-452 (2000).

15. Kun, A., Santos, M. & Szathmary, E. Real ribozymes suggest a relaxed error threshold. *Nat Genet* **37**, 1008-1011 (2005).

16. Rutherford, S.L. & Lindquist, S. Hsp90 as a capacitor for morphological evolution. *Nature* **396**, 336-342 (1998).

17. True, H.L. & Lindquist, S.L. A yeast prion provides a mechanism for genetic variation and phenotypic diversity. *Nature* **407**, 477-483 (2000).

18. Masel, J. Cryptic Genetic Variation Is Enriched for Potential Adaptations. *Genetics* **172**, 1985-1991 (2006).

19. Darwin, C.R. *On the Origin of Species by means of natural selection, or the preservation of favoured races in the struggle for life.* (John Murray: London, 1859).

20. Gould, S.J. & Vrba, E.S. Exaptation; a missing term in the science of form. *Paleobiology* **8**, 4-15 (1982).

21. Pan, J. & Woodson, S.A. Folding intermediates of a self-splicing RNA: mispairing of the catalytic core. *Journal of Molecular Biology* **280**, 597-609 (1998).

22. Rangan, P., Masquida, B., Westhof, E. & Woodson, S.A. Assembly of core helices and rapid tertiary folding of a small bacterial group I ribozyme. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 1574 -1579 (2003).

23. Tanner, M. & Cech, T. Activity and thermostability of the small self-splicing group I intron in the pre-tRNA(lle) of the purple bacterium Azoarcus. *RNA* **2**, 74-83 (1996).

24. Kuo, L.Y., Davidson, L.A. & Pico, S. Characterization of the Azoarcus ribozyme: tight binding to guanosine and substrate by an unusually small group I ribozyme. *Biochimica et Biophysica Acta (BBA) - Gene Structure and Expression* **1489**, 281-292 (1999).

25. Stahley, M.R. & Strobel, S.A. Structural Evidence for a Two-Metal-Ion Mechanism of Group I Intron Splicing. *Science* **309**, 1587-1590 (2005).

26. Beaudry, A.A. & Joyce, G.F. Directed evolution of an RNA enzyme. *Science* **257**, 635-641 (1992).

27. Lehman, N. & Joyce, G.F. Evolution in vitro of an RNA enzyme with altered metal dependence. *Nature* **361**, 182-185 (1993).

28. Blake, R.D. & Delcourt, S.G. Thermodynamic Effects of Formamide on DNA Stability. *Nucleic Acids Research* **24**, 2095 -2103 (1996).

29. Meyers, L.A., Lee, J.F., Cowperthwaite, M. & Ellington, A.D. The Robustness of Naturally and Artificially Selected Nucleic Acid Secondary Structures. *J Mol Evol* **58**, 681-691 (2004).

30. Cadwell, R.C. & Joyce, G.F. Randomization of genes by PCR mutagenesis. *Genome Research* **2**, 28 -33 (1992).

31. O'Brien, P.J. & Herschlag, D. Catalytic promiscuity and the evolution of new enzymatic activities. *Chemistry & Biology* **6**, R91-R105 (1999).

32. Hartl, D.L. & Clark, A.G. *Principles of population genetics*. (Sinauer Associates, Inc.: Sunderland, MA, 1997).

33. Hanczyc, M.M. & Dorit, R.L. Experimental evolution of complexity: in vitro emergence of intermolecular ribozyme interactions. *RNA* **4**, 268-275 (1998).

34. Lawrence, M.S. & Bartel, D.P. New ligase-derived RNA polymerase ribozymes. *RNA* **11**, 1173-1180 (2005).

35. Ouellet, J., Byrne, M. & Lilley, D.M. Formation of an active site in trans by interaction of two complete Varkud Satellite ribozymes. *RNA* **15**, 1822-1826 (2009).

36. Weinreich, D.M., Delaney, N.F., DePristo, M.A. & Hartl, D.L. Darwinian Evolution Can Follow Only Very Few Mutational Paths to Fitter Proteins. *Science* **312**, 111-114 (2006).

37. Kondrashov, F.A. & Kondrashov, A.S. Multidimensional epistasis and the disadvantage of sex. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 12089 -12092 (2001).

38. Wilke, C., Lenski, R. & Adami, C. Compensatory mutations cause excess of antagonistic epistasis in RNA secondary structure folding. *BMC Evolutionary Biology* **3**, 3 (2003).

39. Hermisson, J. & Wagner, G.P. The Population Genetic Theory of Hidden Variation and Genetic Robustness. *Genetics* **168**, 2271-2284 (2004).

40. Aharoni, A. et al. The 'evolvability' of promiscuous protein functions. *Nat Genet* **37**, 73-76 (2005).

41. Bloom, J.D., Labthavikul, S.T., Otey, C.R. & Arnold, F.H. Protein stability promotes evolvability. *Proceedings of the National Academy of Sciences* **103**, 5869 -5874 (2006).

42. Fisher, R. *The Genetical Theory of Natural Selection*. (Oxford University Press, USA: 2000).at <http://www.amazon.ca/exec/obidos/redirect?tag=citeulike09-20&path=ASIN/0198504403>

43. Ferrada, E. & Wagner, A. Protein robustness promotes evolutionary innovations on large evolutionary time-scales. *Proc Biol Sci* **275**, 1595-1602 (2008).

44. Amitai, G., Gupta, R.D. & Tawfik, D.S. Latent evolutionary potentials under the neutral mutational drift of an enzyme. *HFSP J.* **1**, 67-78 (2007).

45. Rydzanicz, R., Zhao, X.S. & Johnson, P.E. Assembly PCR oligo maker: a tool for designing oligodeoxynucleotides for constructing long DNA molecules for RNA production. *Nucleic Acids Research* **33**, W521-W525 (2005).

46. Hochberg, Y. & Benjamini Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B Methodological* **57**, 289-300

Figure 1 | Selection procedure and experimental overview. a, Approximately 10^{13} ribozyme variants (single horizontal black line) are challenged to catalyze the cleavage of an exogenous substrate (single horizontal grey line). Successful ribozymes are specifically amplified by reverse transcription (RT) PCR using a primer (grey arrow) specific to the ligated portion of the substrate (grey line). A second PCR regenerates the active form of the ribozyme using primers (black arrows) specific to its 5' and 3' ends. Transcription of this PCR product generates ribozymes for the next generation. **b**, A diagram of cleavage reaction chemistry. The chemical change (phosphorothioate; PT) for *New* lines is shown by an X at the scissile phosphate. **c,** Experimental overview. The same initial population is used to start lines A (open red) and B (open blue). Cryptic variation is produced by selecting for activity on a native substrate (RNA oligonucleotide) for 10 generations (x10). The populations that have acquired cryptic variation are then used to start two lines, *New*-A (solid red) and *New*-B (solid blue). A third new line, *New*-wt (solid black) is started from the initial population that started A and B. The *New* lines experience an environmental change by requiring selection on a phosphorothioate substrate (b), carried out for 8 generations (x8). The color scheme used here is preserved in subsequent figures.

Figure 2 | Evolution during selection for the native activity. The colors used for each line correspond to those in Fig. 1c. **a,** Activity (fraction ribozyme reacted) at each generation under conditions used during selection for the native activity (RNA oligonucleotide cleavage) over 10 generations (Methods). Error bars correspond to standard errors of three measurments. **b,c,** Histograms, from each generation of line A and B, showing the frequency (percent of sample) of individuals with a given number

of nucleotide differences from the wild-type sequence (distance). Frequencies from generations 1 (G1) and 10 (G10) are shown as solid lines, and intervening generations are shown as dashed grey lines. **d,** Secondary structure of the *Azoarcus* ribozyme according to[25], with nucleotide positions numbered according to[23]. The substrate for selection is shown in lower case letters, with an "x" representing the location of the PT bond used in the *New* lines. Structural elements P4 and J3/4 are indicated. **e,** Representation of positions that accumulated mutations during selection for native activity. The crystal structure of the ribozyme (1ZZN)[25] is shown using the "PUTTY" function in Pymol (Schrödinger). Mutable positions are highlighted by scaling the thickness and color of the tube to the P-value found from linear regression relative to zero (Methods). The active site of the ribozyme is indicated by the substrate (stick representation, colored by element).

 Figure 3 | Evolution during selection for the new activity. a, Activity (fraction ribozyme reacted) at each generation under conditions used during selection for phosphorothioate bond cleavage. Error bars represent standard error of three measurements. **b,** Histograms, from generations of line New-A (red), New-B (blue), and New-wt (black) showing the frequency (percent of sample) of individuals with a given number of nucleotide differences from the wild-type sequence (distance). **c,** Individual mutations that were used to search for rapidly increasing genotypes. The mutations of the *AzoΔ* and *Azo** genotypes are shown in bold. **d,** Frequency of

genotypes, comprised of combinations of mutations in **c**, that showed a rapid increase in frequency (percent of sample) over time (generations), and their corresponding fitness relative to the rest of the population (see text). **e,** Intermolecular activity of the *AzoΔ* ribozyme, under the same conditions as during selection (Methods): 200 pmol

PT substrate, 20 pmol 5'-[^{32}P]-labeled *AzoΔ*. In additon, lanes 3 and 4 contained 40 pmol *wild-type* and *Azo**, respectively. The negative control "No S (–)" contained no substrate. **f,** Comparison of kinetic parameters for the *Azo** and *wild-type* ribozymes.

Figure 4 | Cryptic nature of *Azo** mutations in the native environment. a, Activities (fraction ribozyme reacted) of variants of the *Azoarcus* ribozyme with individual mutations of the *Azo** genotype. Error bars show standard error of at least three measurments. For reference, the error bars of the wild-type activity are extended as grey dashed lines. **b,** Activities (fraction ribozyme reacted) of the wild-type ribozyme and a population of ribozymes generated from random mutagenesis of the *Azo** genotype (μ-Azo*). The activites are shown on either the native substrate (RNA) or the new substrate (PT = phosphorothioate). The RNA reactions were carried out under the same conditions as in line A (no formamide). The PT reactions were carried out under the same conditions as the *New* lines.

Figure 5 | Evolution in genotype space. Principal component analysis of multiple sequence alignments constructed from the combined sequences from New-A, New-B and New-wt populations. Each node represents an individual sequence. The distance between nodes is proportional to the number of nucleotide differences. Note that the apparent distance on the graph may be lower than the true distance due to the compression of multiple dimensions into a 2D plot. Sequences from the indicated generation of each line were combined before alignment. The region on the graphs occupied by the *AzoΔ* sequences is indicated by a grey ellipse.

Figure 6 | Proximity of sequences to the *Azo** genotype. **a,** A network graph composed of all sequences from the last generation of line B (B10), and the first generations of lines *New*-B (*New*-B1) and *New*-wt (*New*-wt1). Nodes represent several sequences clustered at 96% identity (Methods) and are colored based on the source population of their representative sequence (see color key). Edges connect clusters with ≤10 nucleotide differences between their most distant sequences. Nodes containing *Azo** individuals are yellow with a black border. All connections to *Azo** nodes are shown as yellow edges, and the connecting node has a yellow border. The wild-type sequence is a green node with a yellow border. **b-d,** Frequency, presented as percent of sampled sequences (left y-axis) and total number of sequences (right y-axis), of sequences with a given number of the *Azo** mutations in B10, *New*-wt1, and *New*-B1. **e,** Unrooted maximum likelihood phylogenetic tree of sequences found within the yellow nodes of **a** (96% ID with *Azo**) and the representative sequences from *New*-wt1. Leaves represent individual sequences and are colored to match **a** (see key).